Probabilistic methods for multi-source and temporal biomedical data quality assessment

CARLOS SÁEZ SILVESTRE

EDITORIAL
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Probabilistic methods for multi-source and temporal biomedical data quality assessment

Universitat Politècnica de València
Departamento de Ingeniería Electrónica
Programa de Doctorado en
Tecnologías para la Salud y el Bienestar

DOCTORAL THESIS

Presented by
Carlos Sáez Silvestre

Directed by
Dr. Juan M García-Gómez
Dr. Montserrat Robles Viejo

Valencia, Spain
January 2016

Collection Doctoral Thesis

© Carlos Sáez Silvestre

© 2017, of the present edition: Editorial Universitat Politècnica de València
www.lalibreria.upv.es / Ref.: 5838_01_01_01

ISBN: 978-84-9048-646-7

Agradecimientos
Acknowledgements

Esta tesis recopila el resultado de muchos años en los que hacer el trabajo de cada día con dedicación y esfuerzo ha sido la mayor prioridad. No obstante, la efectividad de la dedicación y del esfuerzo se ve claramente beneficiada por un entorno motivador, inspirador y que proporcione los recursos y conocimiento adecuados para avanzar hacia objetivos comunes.

Y es que los actos de una persona en la vida, incluida una tesis doctoral, son reflejo de su experiencia, donde compañeros, amigos y familia contribuyen a configurar quien se es. Por eso quiero aprovechar este espacio para mostrar mi agradecimiento a aquellas personas de quienes he aprendido y que han estado a mi lado en este tiempo. Personas que en mayor o en menor medida han hecho que sea quien soy hoy.

En primer lugar, agradezco a mis directores el Dr. Juan Miguel García-Gómez y la Dra. Montserrat Robles por la confianza común y por el gran apoyo recibido durante más de nueve años, en los que, no solo contando con ellos como directores, sino también como compañeros de investigación, hemos obtenido juntos resultados de los que me siento orgulloso.

El entorno motivador, inspirador, y de conocimiento mencionado anteriormente tiene su componente principal en mis compañeros y amigos de la línea de Minería de Datos Biomédicos del grupo IBIME, con quienes he compartido el día a día y de quienes también he tenido el placer de aprender durante todo este tiempo. Un especial y respetuoso agradecimiento para Miguel Esparza, Salvador Tortajada, Elíes Fuster, Adrián Bresó, Javier Juan, Alfredo Navarro, Alfonso Pérez, Javier Vicente, Juan Martínez y Juan Miguel García-Gómez. Agradezco también por haber contado con su ayuda y por el buen ambiente de trabajo al resto de compañeros de IBIME, VeraTech for Health y el Instituto ITACA, en especial a José Vicente Manjón, José Alberto Maldonado, David Moner, Diego Boscá, José Enrique Romero, Estíbaliz Parcero, Santiago Salas, Crispín Gómez y Vicente Giménez.

I would like to thank Dr. Pedro Pereira Rodrigues for giving me the opportunity to learn from him during the research stay performed at the CINTESIS research group at the University of Porto, with appreciation to the institution head Prof. Altamiro Costa-Pereira. I also thank Dr. João Gama for giving me the opportunity to stay during part of the mobility period in the LIAAD research group. This period contributed to me maturing as a researcher, what also benefited from the technical conversations

and excellent atmosphere with the CINTESIS and LIAAD colleagues and professors, specially Daniel Pereira, Ariane Sasso, Cláudia Camila, Hadi Fanaee, Alexandre Carvalho, Fábio Pinto, Márcia Oliveira, Ricardo Correia, Pavel Bradzil, João Moreira and Carlos Soares.

Quiero agradecer por su motivación científica y colaboración en los casos de estudio de esta tesis a Óscar Zurriaga, Carmen Alberich, Inma Melchor y Jordi Pérez, de la Dirección General de Salud Pública de la Generalitat Valenciana; y a Ricardo García de León González, Ricardo García de León Chocano y Verónica Muñoz del Hospital Virgen del Castillo, Yecla.

This thesis has been developed in the context of several public and private research projects. Therefore, I would like to thank the different funding institutions including the Universitat Politècnica de València; the previous Spanish Ministry of Science and Innovation and current Ministry of Economy and Competitiveness; the Spanish Ministry of Health, Social Services and Equality; the European Commission; Fagor Electrodomésticos S.Coop; IVI Valencia S.L. and VeraTech for Health S.L. I have a special acknowledgement to the excellent people I had the opportunity to work with during my initial research stages in the FP6 European Projects HealthAgents and eTumour. I would also like to give a message of thanks to all the people who put their trust in research, either contributing to public funding, or sharing their data to research studies. My message of hope to the Open Data movement.

Me gustaría cerrar el círculo de agradecimientos dando el mayor de ellos a mi familia. A mis padres Teresa y Vicente, quienes siempre me abrieron todas las puertas al aprendizaje y a la educación. No se puede expresar en palabras el agradecimiento a mi madre por su esfuerzo, que me ha permitido llegar hasta esta etapa. Agradezco además a mis abuelos, mi hermano, y el resto de mi familia por haber estado siempre a mi lado.

Por todo el apoyo recibido durante tanto tiempo, y por su esfuerzo, quiero dedicar esta tesis a Mari, gracias por formar parte de esta historia.

Finalmente, quiero hacer una dedicatoria especial y con todo mi cariño para Claudia, por haberme dado la fuerza necesaria en esta última etapa. Y en las que están por venir.

Abstract

Nowadays, biomedical research and decision making depend to a great extent on the data stored in information systems. As a consequence, a lack of data quality (DQ) may have significant effects in the interpretation of data, which may lead to suboptimal decisions, or hinder the derived research processes and outcomes.

Generally, DQ is assessed by means of evaluating DQ dimensions for fundamental problems such as incomplete, inconsistent or incorrect data. However, the successful development of Big Data and large-scale biomedical repositories, based on multi-institutional, cross-border, data-sharing infrastructures is requiring new approaches for a broad and efficient DQ assessment.

This thesis aims to the research and development of methods for assessing two DQ problems of special importance in large-scale multi-site repositories acquired during long periods of time: (1) the variability of data probability distributions among different data sources or sites—multi-source variability—and (2) the variability of data probability distributions over time—temporal variability. This variability may be caused by differences in data acquisition methods, protocols or health care policies; systematic or random errors during data input and management; geographic and demographic differences in populations; or even falsified data. This variability, if unmanaged, may complicate data analyses, bias the results, or weaken the generalization of hypothesis or models based on the data.

To date, multi-source and temporal variability issues have received little attention as DQ problems nor, to our knowledge, count with adequate assessment methods. This thesis contributes with methods to measure, detect and characterize variability. The methods have been specially designed to overcome the problems that classical statistical approaches may have when dealing with large-scale biomedical data, namely to multi-type, multivariate, multi-modal data, and not affected by large sample sizes such as in Big Data environments. To this end, we have defined an Information Theory and Geometry probabilistic framework supporting the methods. It is based on the inference of non-parametric statistical manifolds from normalized probabilistic distances between distributions among data sources and over time. Based on this probabilistic framework, a number of contributions have been generated.

For the multi-source variability assessment we have designed two metrics: (1) the Global Probabilistic Deviation (GPD), which measures the degree of global variability among the distributions of multiple sources—as an estimator equivalent to the standard deviation among distributions; and (2) the Source Probabilistic Outlyingness (SPO), which measures the dissimilarity of the distribution of a single data source to a global latent average. These metrics are based on the construction of a simplex geometrical figure (the maximum-dimensional statistical manifold) using the distances among data sources. Additionally, we defined Multi-Source Variability (MSV) plot, an exploratory visualization based on that simplex which permits detecting grouping patterns among data sources.

The temporal variability method provides two main tools: (1) the Information Geometric Temporal (IGT) plot, an exploratory visualization of the temporal evolution of data distributions based on the projection of the statistical manifold of relationships among temporal batches; and (2) the Probability Distribution Function Statistical Process Control (PDF-SPC), an algorithm for the monitoring and automatic change detection in data distributions. Additionally, we can monitor the multi-source methods over time.

The methods have been applied to real case studies in biomedical repositories, including: the Public Health Mortality and Cancer Registries of the Region of Valencia, Spain; the UCI Heart Disease dataset; the United States NHDS dataset; a Spanish Breast Cancer dataset; and an In-Vitro Fertilization dataset. A detailed description of the multi-source and temporal variability findings in the Mortality Registry case study is provided, including: a partitioning of the repository into two probabilistically separated temporal subgroups following a change in the Spanish National Death Certificate in 2009, punctual temporal anomalies due to a punctual increment in the number of missing data, along with outlying and clustered health departments due to differences in populations or in practices.

The systematic application of the methods to the case studies has contributed to the development of a software toolbox, which includes the GPD, SPO, MSV plot, IGT plot, PDF-SPC, other basic DQ tools, and the automated generation of DQ reports. Finally, we defined the theoretical basis of a general framework for the evaluation of biomedical DQ, which have been used in three applications: in a process for the construction of quality assured infant feeding repositories, for the contextualization of data for their reuse in Clinical Decision Support Systems using an HL7-CDA wrapper; and in an on-line service for the DQ evaluation and rating of biomedical data repositories.

The results of this thesis have been published in eight scientific contributions, including top-ranked journals and conferences in the areas of Statistics and Probability, Information Systems, Data Mining, Medical Informatics and Biomedical Engineering. One of the journal publications was selected by the IMIA as one of the best publications in 2013 in the subfield of Health Information Systems. Additionally, the results of this thesis have contributed to several research projects, and have facilitated the initial steps towards the industrialization of the developed methods and approaches for the audit and control of biomedical DQ.

Resumen

Actualmente, la investigación y toma de decisiones en entornos biomédicos dependen en gran medida de los datos almacenados en los sistemas de información. En consecuencia, una falta de calidad en los datos (CD) puede afectar significativamente a la interpretación de los mismos, lo cual puede dar lugar a decisiones sub-óptimas o dificultar los procesos y resultados de las investigaciones derivadas.

Generalmente, la CD es evaluada mediante diversas métricas de las denominadas dimensiones de calidad sobre problemas fundamentales como datos incompletos, inconsistentes o incorrectos. Sin embargo, los actuales desarrollos sobre repositorios de datos biomédicos masivos (Big Data), basados en infraestructuras de compartición de datos multi-institucionales o transfronterizas, requieren nuevas aproximaciones para una evaluación eficiente y desde perspectivas generales de la CD.

Esta tesis tiene como propósito la investigación y desarrollo de métodos para evaluar dos problemas especialmente importantes en repositorios multi-sitio masivos adquiridos durante largos periodos de tiempo: (1) la variabilidad de las distribuciones de probabilidad de los datos entre diferentes fuentes o sitios—variabilidad multi-fuente—y (2) la variabilidad de las distribuciones de probabilidad de los datos a lo largo del tiempo—variabilidad temporal. Esta variabilidad puede estar causada por diferencias en los métodos de adquisición de datos, protocolos o políticas de atención sanitaria; a errores sistemáticos o aleatorios durante la entrada o gestión de datos; diferencias geográficas o demográficas en las poblaciones; o incluso por falsificaciones en los datos. Si esta variabilidad no es gestionada, puede complicar el análisis de los datos, sesgar los resultados, o minimizar la generalización de modelos o hipótesis basadas en los datos.

Hasta la fecha, la variabilidad multi-fuente y temporal han recibido poca atención como problemas de CD, y hasta donde sabemos no cuentan con métodos adecuados para su evaluación. Esta tesis aporta métodos para detectar, medir y caracterizar dicha variabilidad, los cuales han sido especialmente diseñados para superar los problemas que las aproximaciones estadísticas clásicas pueden tener con datos biomédicos multi-tipo, multivariantes y multi-modales, y sin ser afectados por tamaños muestrales grandes en entornos Big Data. Para ello, hemos definido un marco probabilístico común basado en Teoría y Geometría de la Información que da soporte a los métodos desarrollados. Este marco está basado en la inferencia de variedades de Riemann no-paramétricas a partir de distancias probabilísticas normalizadas entre las distribuciones de varias fuentes de datos o a lo largo del tiempo. Basadas en dicho marco probabilístico se han aportado las siguientes contribuciones.

Para la evaluación de la variabilidad multi-fuente se han definido dos métricas y un gráfico para visualización: (1) la *Global Probabilistic Deviation* (GPD), la cual mide el grado de variabilidad global entre las distribuciones de las diferentes fuentes—como un estimador equivalente a la desviación estándar entre distribuciones; y (2) la *Source Probabilistic Outlyingness* (SPO), la cual mide la disimilaridad entre la distribución de una fuente de datos dada y la distribución de una fuente promedio o global latente definida. Estas métricas están basadas en la construcción de un simplex geométrico (la variedad de máxima dimensionalidad) mediante las distancias entre fuentes. Adicionalmente, se ha definido el *Multi-Source Variability* (MSV) *plot*, para una visualización exploratoria basada en tal simplex, que permite detectar patrones de agrupamiento o desagrupamiento entre fuentes.

Para la variabilidad temporal el método desarrollado proporciona dos herramientas principales: (1) el *Information Geometric Temporal* (IGT) *plot*, para una visualización exploratoria de la evolución temporal de las distribuciones de datos, basada en la proyección de la variedad estadística de las relaciones entre lotes temporales; y (2) el *Probability Distribution Function Statistical Process Control* (PDF-SPC), un algoritmo para la monitorización y detección automática de cambios en las distribuciones de datos. Adicionalmente, este método permite monitorizar la variabilidad multi-fuente a lo largo del tiempo.

Los métodos han sido aplicados en casos de estudio reales en repositorios biomédicos, incluyendo: el Registro de Salud Pública de Mortalidad y el de Cáncer de la Comunidad Valenciana, España; el conjunto de datos de enfermedades del corazón del repositorio UCI; el conjunto de datos NHDS de los Estados Unidos; un conjunto de datos español de Cáncer de Mama; y un conjunto de datos de Fecundación In-Vitro. En particular esta tesis incluye una descripción detallada de los hallazgos de variabilidad multi-fuente y temporal del Registro de Mortalidad, incluyendo: una partición del repositorio en dos subgrupos temporales probabilísticamente separados siguiendo un cambio en el Certificado Médico de Defunción en 2009, anomalías temporales puntuales debidas a incrementos puntuales en el número de datos perdidos, así como departamentos de salud anómalos y agrupados debido a diferencias en poblaciones y en las prácticas.

La aplicación sistemática de los métodos a los casos de estudio ha contribuido al desarrollo de un conjunto de herramientas software, el cual incluye los métodos GPD, SPO, MSV plot, IGT plot, PDF-SPC, otras herramientas básicas de CD, y la generación automática de informes de CD. Finalmente, se ha definido la base teórica de un marco general de CD biomédicos, el cual ha sido utilizado en tres aplicaciones: en el proceso de construcción de repositorios de calidad asegurada para la alimentación del lactante, en la contextualización de datos para el reuso en Sistemas de Ayuda a la Decisión Médica usando un *wrapper* HL7-CDA, y en un servicio on-line para la evaluación y clasificación de la CD de repositorios biomédicos.

Los resultados de esta tesis han sido publicados en ocho contribuciones científicas (revistas indexadas y artículos en congresos), en las áreas de Estadística y Probabilidad, Sistemas de Información, Minería de Datos, Informática Médica e Ingeniería Biomédica. Una publicación fue seleccionada por la IMIA como una de las mejores publicaciones en 2013 en Sistemas de Información de Salud. Los resultados de esta tesis han contribuido en varios proyectos de investigación, y han facilitado los primeros pasos hacia la industrialización de los métodos y tecnologías desarrolladas para la auditoría y control de la CD biomédica.

Resum

Actualment, la investigació i presa de decisions en entorns biomèdics depenen en gran mesura de les dades emmagatzemades en els sistemes d'informació. En conseqüència, una manca en la qualitat de les dades (QD) pot afectar significativament a la seua interpretació, la qual cosa pot donar lloc a decisions sub-òptimes o dificultar els processos i resultats de les investigacions derivades.

Generalment, la QD és avaluada mitjançant la mesura de dimensions de qualitat sobre problemes fonamentals com dades incompletes, inconsistents o incorrectes. No obstant això, els actuals desenvolupaments sobre repositoris de dades biomèdiques massius (Big Data) basats en infraestructures de compartició de dades multi-institucionals o transfrontereres, requereixen noves aproximacions per a una avaluació eficient i des de perspectives generals de la QD.

Aquesta tesi té com a propòsit la investigació i desenvolupament de mètodes per avaluar dos problemes especialment importants en repositoris multi-lloc, massius i adquirits durant llargs períodes de temps: (1) la variabilitat de les distribucions de probabilitat de les dades entre diferents fonts o llocs—variabilitat multi-font—i (2) la variabilitat de les distribucions de probabilitat de les dades al llarg del temps—variabilitat temporal. Aquesta variabilitat pot estar causada per diferències en els mètodes d'adquisició de dades, protocols o polítiques d'atenció sanitària; a errors sistemàtics o aleatoris durant l'entrada o gestió de dades; diferències geogràfiques o demogràfiques en les poblacions; o fins i tot per falsificacions en les dades. Si aquesta variabilitat no és gestionada, pot complicar l'anàlisi de les dades, esbiaixar els resultats, o minimitzar la generalització de models o hipòtesis basades en les dades.

Fins a la data, la variabilitat multi-font i temporal han rebut poca atenció com problemes de QD, i fins on sabem no compten amb mètodes adequats per a la seva avaluació. Aquesta tesi aporta mètodes per detectar, mesurar i caracteritzar aquesta variabilitat. Aquests mètodes han estat especialment dissenyats per superar els problemes que les aproximacions estadístiques clàssiques poden tenir amb dades biomèdiques multi-tipus, multivariants i multi-modals, i per a no ser afectats per mides mostrals grans en entorns Big Data. Per a això, hem definit un marc probabilístic comú basat en Teoria i Geometria de la Informació que dóna suport als mètodes desenvolupats. Aquest marc està basat en la inferència de varietats de Riemann no-paramètriques a partir de distàncies probabilístiques normalitzades entre les distribucions de diverses fonts de dades o al llarg del temps.

Basades en aquest marc probabilístic s'han aportat les següents contribucions:

Per a l'avaluació de la variabilitat multi-font s'han definit dos mètriques i un gràfic per a visualització: (1) la Global Probabilistic Deviation (GPD), la qual mesura el grau de variabilitat global entre les distribucions de les diferents fonts—com un estimador equivalent a la desviació estàndard entre distribucions; i (2) la Source Probabilistic Outlyingness (SPO), la qual mesura la dissimilaritat entre la distribució d'una font de dades donada i la distribució d'una font mitjana o global latent definida. Aquestes mètriques estan basades en la construcció d'un simplex geomètric (la varietat de màxima dimensionalitat) mitjançant les distàncies entre fonts. Addicionalment, s'ha definit el Multi-Source Variability (MSV) plot, per a una visualització exploratòria basada en tal simplex, que permet detectar patrons d'agrupament o desagrupament entre fonts.

Per a la variabilitat temporal el mètode desenvolupat proporciona dues eines principals: (1) l'Information Geometric Temporal (IGT) plot, per a una visualització exploratòria de l'evolució temporal de les distribucions de dades, basada en la projecció de la varietat estadística de les relacions entre lots temporals; i (2) el Probability Distribution Function Statistical Process Control (PDF-SPC), un algoritme per al monitoratge i detecció automàtica de canvis en les distribucions de dades. Addicionalment, aquest mètode permet monitoritzar la variabilitat multi-font al llarg del temps.

Els mètodes han estat aplicats en casos d'estudi reals en repositoris biomèdics, incloent: el Registre de Salut Pública de Mortalitat i el de Càncer de la Comunitat Valenciana, Espanya; el conjunt de dades de malalties del cor del repositori UCI; el conjunt de dades NHDS dels Estats Units; un conjunt de dades espanyol de Càncer de Mama; i un conjunt de dades de Fecundació In-Vitro. En particular la tesi inclou una descripció detallada de les troballes de variabilitat multi-font i temporal del Registre de Mortalitat, incloent: una partició del repositori en dos subgrups temporals probabilísticament separats seguint un canvi en el Certificat Mèdic de Defunció el 2009, anomalies temporals puntuals degudes a increments puntuals en el nombre de dades perdudes, així com departaments de salut anòmals i agrupats a causa de diferències en poblacions i en les pràctiques.

L'aplicació sistemàtica dels mètodes als casos d'estudi ha contribuït al desenvolupament d'un conjunt d'eines programari, el qual inclou els mètodes GPD, SPO, MSV plot, IGT plot, PDF-SPC, altres eines bàsiques de QD, i la generació automàtica d'informes de QD. Finalment, s'ha definit la base teòrica d'un marc general de QD biomèdiques, el qual ha estat utilitzat en tres aplicacions: en el procés de construcció de repositoris de qualitat assegurada per l'alimentació del lactant, a la contextualització de dades per a la reutilització en Sistemes d'Ajuda a la Decisió Mèdica usant un wrapper HL7-CDA, i en un servei on-line per a l'avaluació i classificació de la QD de repositoris biomèdics.

Els resultats d'aquesta tesi han estat publicats en vuit contribucions científiques (en publicacions en revistes indexades i en articles en congressos), en les àrees d'Estadística i Probabilitat, Sistemes d'Informació, Mineria de Dades, Informàtica Mèdica i Enginyeria Biomèdica. Una de les publicacions va ser seleccionada per la IMIA com una de les millors publicacions en 2013 en la sub-àrea de Sistemes d'Informació de Salut. Addicionalment, els resultats d'aquesta tesi han contribuït en diversos projectes d'investigació, i han facilitat les primeres passes cap a la industrialització dels mètodes i aproximacions desenvolupades per l'auditoria i control de la QD biomèdica.

Glossary

Mathematical notation

x	Random variable
$p(x)$	Probability (density/mass) function of a variable X
$p(x, y)$	Joint probability function of two random variables X and Y
$\mathbf{\Theta}$	Vector of parameters for a probability function
$p(x\|\mathbf{\Theta})$	Probability function of a variable X conditioned to a vector of parameters $\mathbf{\Theta}$
$D(P\|\|Q)$	Dissimilarity between probability distributions $P = p(x)$ and $Q = q(x)$
$D_z(X, Y)$	Dissimilarity between elements X and Y under the metric conditions of z
$d(X, Y)$	Distance between elements Y and Y
$\mathcal{M}$	Riemannian manifold
$\mathbf{x}$	Column vector $\mathbf{x}$
$\mathbf{x}^T$	Transpose of $\mathbf{x}$
$\mathbb{E}p(x)$	Expected value of probability distribution $p(x)$
$\mathbb{R}^D$	D-dimensional space of real numbers
$\mathbb{N}$	Space of natural numbers
$\frac{\partial x}{\partial t}$	Partial derivative of variable x with respect to variable t
∇	Nabla operator
$\|\| \cdot \|\|$	Euclidean norm
Δ^D	D-dimensional simplex geometric figure
$d_{1R}(D)$	Maximum possible distance between any vertex and the centroid in a D-dimensional regular simplex with edge length of 1
$d_{max}(D)$	Maximum possible distance between any vertex and the centroid in a D-dimensional irregular simplex
Ω	Symbol for Global Probabilistic Deviation
$\mathbb{O}$	Symbol for Source Probabilistic Outlyingness
$\{\cdot\}$	Set of elements

Acronyms

ANOVA Analysis of Variance

BHFI Baby-friendly Hospital Initiative

BMI Body Mass Index

CRRV Cancer Registry of the Region of Valencia

CDF Cumulative Density Function

CDSS Clinical Decision Support System

CONSORT Consolidated Standard of Reporting Trials

CSV Comma-Separated Values

DQ Data Quality

EHR Electronic Health Record

EM Expectation Maximization

EMD Earth Mover's Distance

FDA Functional Data Analysis

FIM Fisher Information Matrix

GPD Global Probabilistic Deviation

GUI Graphical User Interface

HIS Health Information System

HL7-CDA Health Level 7 Clinical Document Architecture

ICD International Classification of Diseases

IBIME Biomedical Informatics Group

ID Identifier

IF Impact Factor

IGT Information Geometric Temporal

IMIA International Medical Informatics Association

ITACA Institute of Information and Communication Technologies

IVF In-Vitro Fertilization

JCR Journal Citation Reports

JF Jeffrey Divergence

JS Jensen-Shannon Divergence

JSD Jensen-Shannon Distance

KDE Kernel Density Estimation

KL Kullback-Leibler Divergence

MDS Multidimensional Scaling

MLE Maximum-Likelihood Estimation

MR Magnetic Resonance

MRRV Mortality Registry of the Region of Valencia

MRS Magnetic Resonance Spectroscopy

MSV Multi-Source Variability

NHDS National Hospital Discharge Survey

OLAP On-Line Analytical Processing

PCA Principal Component Analysis

PDF Probability Distribution Function

PDF-SPC Probabilistic Statistical Process Control

SNOMED-CT Systematized Nomenclature of Medicine - Clinical Terms

SPC Statistical Process Control

SPO Source Probabilistic Outlyingness

SV Single Voxel

TDQM Total Data Quality Management

TQM Total Quality Management

UCI University of California, Irvine

UPV Universitat Politècnica de València

US United States

WHO World Health Organization

XML eXtensible Markup Language

Contents

Chapter 1

Introduction

This chapter presents the outline of the thesis. Its main motivations are introduced in first place. These lead to the definition of the thesis research questions and objectives, described in second place. Third, the contributions derived from the research carried out in this thesis are described. Next the projects and partners which have established the work context of this thesis are compiled. Finally, an outline of the thesis structure is provided.

1.1 Motivation

The Biomedical Informatics Group (IBIME) of the Institute of Information and Communication Technologies (ITACA) of the Universitat Politècnica de València (UPV) was established in 1999 as an interdisciplinary research group committed to biomedical informatics. Since then, IBIME has participated in many research activities aimed to the knowledge discovery and modelling from biomedical data.

The author of this thesis joined IBIME in 2006 with a University grant while completing his Master studies. The author initially focused to investigate and develop new Clinical Decision Support Systems (CDSSs) for brain tumour diagnosis, in the framework of the European research projects eTumour and HealthAgents. Further research until the development of this thesis was focused to improve CDSSs machine learning technologies as well as evaluate their effectiveness on real clinical uses.

Given the experience gained by IBIME and the author during those years, a turning point was achieved. Most CDSSs research in IBIME was aimed to systems which knowledge is acquired from empirical, data mining analyses of biomedical data repositories. In such a way, a common issue was found which hindered these investigations. Such an issue consisted in Data Quality (DQ) problems.

Of course, the data quality problem was not exclusive to IBIME's research. At the time of a successful establishment of Electronic Health Records (EHRs), the development of biomedical data sharing technologies and predictive analytics were moving medical informatics to the Big Data era. As a consequence, in the coming age of data-driven, precision medicine, the problematic of DQ for data reuse was becoming more evident, which was reflected in the most relevant international medical informatics journals and conferences.

However, most data quality assessment approaches seemed to give little attention to two problems which will gain importance as larger multi-source repositories are established. These problems are (1) the variability of data probability distributions among multiple sources and (2) the variability of data probability distributions through time. Both problems were already faced in our investigations, and were supported by most recent machine learning research. However, few solutions existed for their assessment as part of DQ procedures, nor were suitable to common characteristics of biomedical data, namely multiple types of variables, with multiple distribution modes, and in a multi-variate setting. In addition, newer solutions should be adapted to Big Data environments, providing scalable methods suitable to large sample sizes.

The aforementioned problems established the main motivations of this thesis, and led to the following research questions and aims.

1.2 Research questions and objectives

A successful data reuse depends in great measure on the quality of its data. Independently of the objective of the data reuse, such as deriving hypotheses or building statistical models, when using a set of biomedical data which has been acquired from distinct sources and/or through a significant period of time, two problems may arise. First, the probability distributions of data may not be concordant among the multiple sources. In some situations, this may be due to expected population differences, however, in others it may be related to unexpected differences or biases in the original samples or in the data acquisition processes—e.g., different protocols for patient data acquisition at the different hospitals. And second, the probability distributions of data may not be concordant through the time period data were acquired. Similarly, this may be due to changes in the processes generating data—e.g., a normative change of protocol—or in the original sources of information—e.g, an environmental change.

When *multi-source variability* and *temporal variability* are not considered for data reuse they may lead to different problems, such as suboptimal data analytics processes, biased or ungeneralizable research results or even inaccurate strategic and healthcare decisions. This thesis was conceived with the purpose of help addressing these problems, and facilitate the data reuse from valid a reliable information. As a consequence, the next research questions arised:

RQ1 To what extent current data quality methods consider the problems of variability of data distributions among multiple sources and through time?

RQ2 Can we provide a metric measurement of the probabilistic variability of data among multiple sources?

RQ3 Can we detect, measure and characterize changes in the probability distributions of biomedical data through time?

RQ4 Will multi-source and temporal variability assessment methods be robust to data with multiple types of variables, multi-modal, multi-variate data and independent to sample size in Big Data environments?

RQ5 Can multi-source and temporal variability assessment methods be part of general data quality assessment procedures for biomedical data?

The research work carried out in this thesis tries to answer these questions, while aiming to define and build empirically driven and validated solutions to address the problems of multi-source and temporal variability in a data quality assessment context. To this end, the next objectives were defined:

O1 Review the state-of-the-art about data quality assessment methods, with a special focus on solutions for assessing and measuring the multi-source and temporal variability.

O2 Evaluate the feasibility of statistical and information theoretic methods as assessment methods and metrics for multi-source and temporal variability.

O3 Design and build a method for the assessment of multi-source variability of biomedical data, specially focusing to provide a robust metric of probabilistic variability.

O4 Design and build a method for the assessment of temporal variability of biomedical data, which facilitates detecting, measuring and characterizing changes.

O5 Validate the methods to be built both on simulated benchmarks and on real biomedical data.

O6 Define the foundations of a framework for the generic assessment of biomedical data quality, which considers the systematic assessment of multi-source and temporal variability dimensions.

The aforementioned objectives are put in common into the final aim of this thesis: *to improve the effectiveness and efficiency of the reuse of biomedical data by means of a reliable information about their quality.* Although it is difficult to provide a global level of measurement for such an aim, it can be decomposed by the accomplishment of the thesis objectives, which are supported by the resultant thesis scientific contributions described next.

1.3 Thesis contributions

This thesis has led to different scientific contributions and technological results. The contributions originated from this thesis work are listed next according to their type. Additionally, the last point of the section provides a brief summary of the author's previous contributions which established the background knowledge and motivation for this thesis.

1.3.1 Main contributions

The main contributions of this thesis are summarized as follows:

C1 - Comparative study of probability distribution distances

This contribution consists in a comparative review of statistical and Information-Theoretic methods for measuring distances between probability distributions. The comparison describes the capabilities of different methods to deal with multi-modal, multi-type and multivariate data, whether they can be bounded, and shows comparison charts of their normalized response on different simulated distribution dissimilarities. This work was published in the conference contribution P2 (Sáez et al, 2013b), and helped to define the scientific basis for the methods in contributions C2 and C3.

C2 - Methods for multi-source variability assessment

Two metrics are proposed. The first is the **Global Probabilistic Deviation (GPD)—C2.1—**, which provides a bounded degree of the global multi-source variability, as an estimator of the probabilistic standard deviation among the different sources. The second is the **Source Probabilistic Outlyingness (SPO)—C2.2—**, which provides a bounded degree of the dissimilarity of each source to a latent central Probability Distribution Function (PDF). The metrics are based on the projection of a simplex geometrical structure constructed from the pairwise probabilistic distance among the sources, represented by the vertices. Besides, the simplex centroid represents a latent central PDF, avoiding the need of a gold standard reference dataset. Additionally, the 2D (or 3D) simplicial projection of the simplex can be used as a visualization method of the multi-source variability, namely the the **Multi-Source Variability (MSV) plot—C2.3—**, which permits exploring any outlying or grouping behaviour of sources. These methods were published in the journal contribution P3 (Sáez et al, 2014b), and compiled in the software contribution S1.

C3 - Methods for temporal variability assessment

A set of automatic and exploratory methods for assessing the variability of biomedical data through time are proposed. The first is the **Information Geometric Temporal (IGT) plot—C3.1—**, a method to analyse the temporal behaviour of data which permits detecting, measuring and characterizing temporal changes in distributions. It relies on a non-parametric information-geometric statistical manifold, which points represent the PDF of consecutive time batches laid out maintaining the probabilistic distance to each other. Two overlapping points would indicate exact PDFs, while a normed distance of 1 between them would indicate completely disjoint PDFs. IGT plots help discovering data temporal trends, conceptually-related time periods, abrupt changes and punctual anomalies. The second is the **Probabilistic Statistical Process Control (PDF-SPC)—C3.2—**, a non-parametric statistical process control for monitoring changes in data distributions through time. Warning and Out-of-Control states are reached according to statistical thresholds (e.g., based on the three-sigma rule) on the Beta distribution of accumulated PDF distances to a moving reference distribution. Out-of-control states confirm new data concepts and re-establish the reference distribution. Finally, the combined use of temporal and multi-source methods provides an information geometric temporal monitoring of multiple sources—**C3.3**. These methods were published in the journal contribution P4 (Sáez et al, 2015), and provided in the software contribution S1.

C4 - Data Quality Assessment reports on real case studies
A set of DQ Assessment reports have been provided for the real case studies on which the methods of this thesis have been validated. These include the Spatio-Temporal Data Quality Assessment of the Mortality Registry of the Comunitat Valenciana—**C4.1**—; Spatio-Temporal Data Quality Assessment of the Cancer Registry of the Comunitat Valenciana—**C4.2**—; Data Quality and Preparation Report for a Sentinel Node Biopsy predictive model in Breast Cancer—**C4.3**—; and Data Quality and Preparation Report for a Twin-Pregnancy Risk Prediction Model with Oocyte Donation—**C4.4**. The results of the contribution C4.1 have been accepted in the journal contribution P5 (Sáez et al, 2016).

C5 - Multi-source and temporal variability software
The data quality methods designed and validated in this thesis for multi-source and temporal variability have been compiled into a software to facilitate its systematic use and as a preparation for a further industrialization. This set of tools include the multi-source and temporal methods, but also other methods were developed to assess missing data, outlier-based inconsistencies and variable predictive value. Additionally, an automatic report generation system was built which automatically constructs a LaTeX-based document with the corresponding data quality results and figures. The software was registered in the technological offer of the Universitat Politècnica de València, as shown in contribution S1. Additionally, the proposal of a systematic use of this software was applied to the case study in contribution C4.1, being as well under review in the journal contribution P5 (Sáez et al, 2016).

C6 - Proposal of a generic Data Quality Assessment framework
Supported by the knowledge about DQ acquired during the development of the thesis, we proposed the definition of a theoretical framework for biomedical DQ assessment—**C6.1**. This framework is based on the definition of nine DQ dimensions aiming to cover the most important dimensions to our opinion, while including the new multi-source and temporal methods and dimensions. Dimensions can be measured in different axes of the dataset, namely through registries, attributes, single-values, full dataset, multi-source and through time. With this contribution we aimed to provide insights into further research in other DQ dimensions alone or in combination with the multi-source and temporal variability problems, towards the application and industrialization of a general DQ framework. Therefore, during the development of this thesis the contents of this framework were used in three applications: (1) to establish the theoretical basis of a process for the construction of quality assured infant feeding repositories—**C6.2**—, (2) in the contextualization of data for its reuse in rule-based CDSS using an HL7-CDA wrapper—**C6.3**—, and to establish the measurements of

DQ features in an on-line service for the evaluation and rating of biomedical data repositories—**C6.4**. The original ideas of this framework were published in the conference contribution P1. The contribution C6.2 has been published in publications P8 (García de León Chocano et al, 2015) and P9 (García de León Chocano et al, 2016). The contribution C6.3 was published in publication P6 (Sáez et al, 2013a) and is provided in the software contribution S2. Finally, the contribution C6.4 is currently in industrialization in a joint partnership of the IBIME research group and the spin-off of the UPV Veratech for Health S.L., provided in the software contribution S3.

The ideas developed during this thesis permitted obtaining funds from the National Government towards the industrialization of the general DQ framework, including the methods for multi-source and temporal variability assessment developed in this thesis, and to perform the necessary additional research for completing our approach. Hence, in addition to the scientific and technological contributions, this thesis has directly led to the creation of job positions and new research projects.

1.3.2 Scientific publications

The contributions of this thesis have led to six journal publications, two conference papers and one book chapter. The journal publications describe the methods for multi-source and temporal variability—P3 and P4—, the application to the Mortality Registry—P5—, the application of the DQ framework for the data contextualization for data reuse by CDSSs—P6—, and the application of the DQ framework to the extraction of quality assured perinatal repositories—P8 and P9. The journals on which this thesis has contributed are top-ranked in the areas of Information Systems, Statistics and Probability, Data Mining and Medical Informatics, according to the Impact Factor (IF) of the Journal Citation Reports (JCR) by Thomson Reuters.

Regarding to the conference papers, the first stands as a position paper resulted after the initial state-of-the-art review—P1— establishing some general concepts for DQ assessment, and the second disseminated the results of the comparative study of probability distribution distances—P2. The two conferences are relevant international scientific conferences on Medical Informatics and Biomedical Engineering.

Finally, the work in the Public Health Mortality and Cancer Registries resulted in an invitation to write two chapter sections related to Data Quality in a guideline for the governance of Public Health patient registries by the PARENT European Project—P7.

The publications of this thesis are listed as follows:

P1 - **Carlos Sáez**, Juan Martínez-Miranda, Montserrat Robles and Juan M García-Gómez. *'Organizing data quality assessment of shifting biomedical data'*. Studies in Health Technology and Informatics, Proceedings of the 24th Medical Informatics in Europe Conference (MIE2012); 180:721-725. Pisa, Italy. August 2012 (Sáez et al, 2012b).

P2 - **Carlos Sáez**, Montserrat Robles and Juan M García-Gómez. *'Comparative study of probability distribution distances to define a metric for the stability of multi-source biomedical research data'*. Proceedings of the 35th annual international conference of the IEEE Engineering in medicine and biology society (EMBC), 3226–3229. Osaka, Japan. July 2013 (Sáez et al, 2013b).

P3 - **Carlos Sáez**, Montserrat Robles and Juan M García-Gómez. *'Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances'.* Statistical Methods in Medical Research. Published Online First in August 2014 (Sáez et al, 2014b).

IF: 4.472 (JCR 2014): 1/122 Statistics and Probability (Q1), 1/24 Medical informatics (Q1), 3/89 Health Care Sciences and Services (Q1), 5/56 Mathematical and Computational Biology (Q1)

P4 - **Carlos Sáez**, Pedro Pereira Rodrigues, João Gama, Montserrat Robles and Juan M García-Gómez. *'Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality'.* Data Mining and Knowledge Discovery. 29(4):950–75. July 2015 (Sáez et al, 2015).

IF: 1.987 (JCR 2014): 25/139 Computer Science, Information Systems (Q1), 41/123 Computer Science, Artificial Intelligence (Q2)

P5 - **Carlos Sáez**, Oscar Zurriaga, Jordi Pérez-Panadés, Inma Melchor, Montserrat Robles and Juan M García-Gómez. *'Applying probabilistic temporal and multi-site data quality control methods to a public health mortality registry in Spain: A systematic approach to quality control of repositories'.* Accepted in the Journal of the American Medical Informatics Association (Sáez et al, 2016).

IF: 3.504 (JCR 2014): 2/24 Medical Informatics (Q1), 6/89 Health Care Sciences and Services (Q1), 8/139 Computer Science, Information Systems (Q1), 9/102 Computer Science, Interdisciplinary Applications (Q1)

P6 - **Carlos Sáez**, Adrián Bresó, Javier Vicente, Montserrat Robles and Juan M García-Gómez. *'An HL7-CDA wrapper to facilitate the semantic interoperability to rule-based clinical decision support systems'.* Computer Methods and Programs in Biomedicine. 109(3):239-249. March 2013 (Sáez et al, 2013a).

Selected as 'Best of medical informatics papers published in 2013, sub-field of Health Information Systems' by the International Medical Informatics Association (IMIA), in the IMIA Yearbook 2014 (Toubiana and Cuggia, 2014).

IF: 1.093 (JCR 2013): 32/102 Computer Science, Theory and Methods (Q2), 16/24 Medical informatics (Q3), 54/76 Engineering, Biomedical (Q3), 68/102 Computer Science, Interdisciplinary Applications (Q3)

P7 - Óscar Zurriaga, Carmen López Briones, Miguel A. Martínez-Beneito, Clara Cavero-Carbonell, Rubén Amorós, Juan M. Signes, Alberto Amador, **Carlos Sáez**, Montserrat Robles, Juan M. García-Gómez, Carmen Navarro-Sánchez, María J. Sánchez-Pérez, Joan L. Vives-Corrons, María M. Mañú, Laura Olaya. *'Methodological guidelines and recommendations for efficient and rational governance of patient registries. Chapter 8: Running a Registry'.* National Institute of Public Health, Trubarjeva 2, 1000 Ljubljana, Slovenia. ISBN:978-961-6911-75-7(pdf) (Zurriaga et al, 2015).

P8 - Ricardo García de León Chocano, **Carlos Sáez**, Verónica Muñoz-Soler, Ricardo García de León González and Juan M García-Gómez. *'Construction of quality-assured infant feeding process of care data repositories: definition and design (Part 1)'.* Computers in Biology and Medicine. 67:95-103. December 2015 (García de León Chocano et al, 2015).

IF: 1.240 (JCR 2014): 49/85 Biology (Q3), 64/102 Computer Science, Interdisciplinary Applications (Q3), 52/76 Engineering, Biomedical (Q3), 35/56 Mathematical and Computational Biology (Q3)

P9 - Ricardo García de León Chocano, Verónica Muñoz-Soler, **Carlos Sáez**, Ricardo García de León González and Juan M García-Gómez. *'Construction of quality-assured infant feeding process of care data repositories: construction of the perinatal repository (Part 2)'.* Accepted in Computers in Biology and Medicine. (García de León Chocano et al, 2016).

1.3.3 Software

The research carried out in this thesis has led to three software developments. First, the methods for multi-source and temporal variability DQ assessment developed in this thesis were compiled in a software toolbox, registered in the software registry of the UPV as part of the University technological offer. Second, the proposed DQ framework was considered in a software to assure the 'contextualization' of patient data for its reuse in rule-based CDSSs. And third, the methods for multi-source and temporal variability assessment and the proposed DQ framework have been included in the design of an industrial application by a partnership formed by the IBIME research group and the company VeraTech for Health S.L., a technological start-up formed by some members of IBIME (including the author and advisers) and spin-off of the UPV.

S1 - **Carlos Sáez**, Juan M García-Gómez, Montserrat Robles and Miguel Esparza. *'R-16880-2014 - Evaluación y rating de la calidad de repositorios de datos biomédicos (DQV)'.* CARTA Registry of the Universitat Politècnica de València.

S2 - **Carlos Sáez**, Adrián Bresó, Javier Vicente, Montserrat Robles and Juan M García-Gómez. *'HL7-CDA Wrapper for the contextualization of biomedical data for reuse in rule-based CDSSs'.*

S3 - IBIME (UPV) and VeraTech for Health S.L. *'Qualize: Quality evaluation and rating of biomedical data repositories'.* Funded by the Spanish Ministry of Economy and Competitiveness (Retos-Colaboración 2013 programme, RTC-2014-1530-1, 2013-2016)

1.3.4 Other contributions

The background knowledge and motivation for this thesis not only arised from the state-of-the-art requirements, but also from own experience of the author and advisors in reusing biomedical data for CDSSs. Such work was done in the framework of several European and National projects, described in the next section, and several scientific and technological contributions were originated from it.

The first group of contributions relate to the research in CDSSs for brain tumour diagnosis carried out within the European projects eTUMOUR and HealthAgents, which culminated in the generic machine learning-based CDSS CURIAM and its specialization for brain tumour diagnosis based on Magnetic Resonance Spectroscopy (MRS) data CURIAM BT (registered in the UPV software registry nos. R13391-2009 and R13392-2009). Hence, the author work in such research and development originated two conference contributions (Sáez et al, 2008, 2009) and one journal publication (Sáez et al, 2011) as main author, and collaborated and co-authored four conference papers (García–Gómez et al, 2007; Croitoru et al, 2007; Xiao et al, 2007, 2008), two journal publications (Hu et al, 2011; Fuster-Garcia et al, 2011), one book chapter (Lluch-Ariet et al, 2007), and several Deliverables of the European projects. The generic capabilities of the CDSS CURIAM led, in addition to its brain tumour specialization, to other specific CDSSs for soft tissue tumours and post-partum depression (Sáez et al, 2008), as well as to a paediatric specific CURIAM BT version (Vicente et al, 2012). Further, the author carried out a randomized pilot study and qualitative evaluation of CURIAM BT in three hospitals in the Region of Valencia: Hospital Universitario Dr. Peset, Hospital de La Ribera and Hospital Quirón Valencia, which originated one conference (Sáez et al, 2012a) and one journal publication (Sáez et al, 2014a), and led CURIAM BT to obtain the award 'Best Technology and Research contribution' in 2012 by the Spanish healthcare Editorial Company 'SANITARIA 2000'. In addition, the work by the author, advisors and colleagues regarding to magnetic resonance in the network led by Dr. Luis Martí-Bonmatí was awarded by the 'Exemplary group in science and academic life: PRO ACADEMIA PRIZE 2013'.

In parallel to that work, the author participated in other research and industrial projects with the following contributions: an evaluation of the user acceptance of a new Health Information System (HIS) in the Balearic Islands for the regional Government; the development of automatic classification modules for a CDSS in ophthalmology; and the data preparation, understanding and quality assessment for a twin-pregnancy risk prediction model in oocyte donation programme, with a journal paper under review. The author additionally participated in the European project HELP4MOOD, contributing in the design and development of a Knowledge Extraction and Inference engine for the management and care of patients with major depression, as well as writing the corresponding Deliverable.

Finally, the author participated actively in a private project aiming to a knowledge-based personal health system for the empowerment of outpatients with diabetes mellitus (Sáez et al, 2013a; Bresó et al, 2015). The requirements of a high-quality patient data reuse for such a knowledge-based CDSS established the start point of this thesis, which approach based on patient and CDSS results standardization originated the first journal contribution JC1, awarded by the International Medical Informatics Association in 2014.

1.4 Projects and partners

From the thesis antecedents to the development of the thesis work, the author has been actively involved in several European, National, private and University-funded projects, collaborating with clinical, academic and private sector partners.

The projects mainly related with the development of this thesis are listed as follows:

DQV-AUTOPROJECT *Servicio de evaluación y rating de la calidad de repositorios de datos biomédicos.* Funded by own IBIME funds - Universitat Politècnica de València (2013-2014)

Objectives: This project aims to an holistic data quality assessment based on the definition of a data quality system by which institutions may evaluate and compare the quality of their datasets.

Partners: IBIME-ITACA group of the Universitat Politècnica de Valencia (Valencia, Spain)

DQV-MINECO *Servicio de evaluación y rating de la calidad de repositorios de datos biomédicos.* Funded by the Spanish Ministry of Economy and Competitiveness (Retos-Colaboración 2013 programme, RTC-2014-1530-1, 2013-2016). **Objectives**: This project aims to define a data quality evaluation and rating service to assure the data value aimed to its reuse in clinical, strategic and scientific decision making. It will be based on two software services. The first will evaluate nine data quality dimensions. The second will generate a data quality rating positioning the evaluated datasets according to several reuse knowledge extraction purposes.

Partners: VeraTech for Health S.L. (Valencia, Spain) and IBIME-ITACA group of the Universitat Politècnica de Valencia, (Spain)

DQV-SPATIO-TEMPORAL-Evaluation *Servicio de evaluación de la estabilidad espacio temporal de repositorios de datos biomédicos.* Funded by the Universitat Politècnica de València (Prueba de Concepto 2015, SP20141432, 2014-2015).

Objectives: The objective of this project is to construct a proof of concept of a spatio-temporal data quality assessment methodology. Concretely, the project aims to access data repositories from reputed centres and generate data quality reports which will reflect the data variability problems and the recommendations to improve their data acquisition and reuse processes.

Partners: Dirección General de Salud Pública, Generalitat Valenciana (Valencia, Spain) and IBIME-ITACA group of the Universitat Politècnica de Valencia (Valencia, Spain)

The projects on which the author was actively involved in parallel the development of this thesis, and previous to it establishing the thesis basis are listed as follows:

eTUMOUR *Web accessible Magnetic Resonance (MR) decision support system for brain tumour diagnosis and prognosis, incorporating in vivo and ex vivo genomic and metabolomic data.* Funded by the European Commission (VI Framework Program, LSHC-CT-2004-503094, 2004-2009).

Objectives: (1) Development of a web-accessible CDSS that has a Graphical User Interface (GUI) to display clinical, metabolomic and genetic brain tumor data. (2) To provide an evidence-based clinical decision-making computer-human interface by using statistical pattern recognition analysis of molecular images of brain tumours (using MRS) and incorporating new criteria such as genetic based tumour classifications and related clinical information.

Partners: University of Valencia (Valencia, Spain), Universitat Autò-noma de Barcelona (Barcelona, Spain), St George's Hospital Medical School (London, UK), University Medical Center Nijmegen (Nijmegen, Netherlands), Stichting Katholieke Universiteit (Nijmegen, Netherlands), Université Joseph Fourier U594 (Grenoble, France), MicroArt S.L. (Barcelona, Spain), Hospital San Joan de Deu (Esplugues de Llobregat, Spain), Pharma Quality Europe, s.r.l. (Barcelona, Spain), Hyperphar Group SpA. (Milan, Italy), Katholieke Universiteit Leuven (Leuven, Belgium), Siemens AG, Medical Solutions (Erlangen, Germany), SCITO, S.A (Grenoble, France), Deutsche Krebsforschungs zentrum Heidelberg (Heidelberg, Germany), Bruker Biospin SA. (Wissembourg, France), Institute of Child Health - University of Birmingham (Birmingham, United Kingdom), INSERM U318 (Grenoble, France), Fundación para la Lucha contra Enfermedades Neurológicas de la Infancia (Buenos Aires, Argentina), Medical University Lodz (Lodz, Poland) and IBIME-ITACA group of the Universitat Politècnica de Valencia (Valencia, Spain).

HEALTHAGENTS *Agent-based distributed decision support system for brain tumour diagnosis and prognosis.* Funded by the European Commission (VI Framework Program, IST-2004-27214, 2006-2009).

Objectives: To create a distributed datawarehouse with the world's largest network of interconnected databases of clinical, histological, and molecular phenotype data of brain tumour patients, providing evidence-based clinical decision-making by means of magnetic resonance and genetic based tumour classifications, and to develop new methodologies to fulfill a dynamic clinical decision support system.

Partners: University of Valencia (Valencia, Spain), MicroArt S.L. (Barcelona, Spain), Universitat Autònoma de Barcelona (Barcelona, Spain), Pharma Quality Europe, s.r.l. (Barcelona, Spain), Katholieke Universiteit Leuven (Leuven, Belgium), University of Birmingham (Birmingham, UK), University of Edinburgh (Edinburg, UK), University of Southampton (Southampton, UK) and IBIME-ITACA group of the Universitat Politècnica de Valencia (Valencia, Spain)

RISCO *Servicio remoto de atención sanitaria basado en la prevención, autonomía y autocontrol de los pacientes.* Funded by the Spanish Ministry of Science and Innovation (INNPACTO 2011, 2011-2013)

Objectives: The objective of the consortium members is being introduced in the international market of Remote Services for Healthcare and Wellbeing based on the development and validation of a technological platform for a novel healthcare service based in the prevention, autonomy and self-control of patients. Concretely, the platform will be aimed to diabetes mellitus and cardiovascular diseases patients, developing novel solutions for patient telemonitoring, risk assessment, nutrition and physical exercise planning, with a multichannel remote assistance specialized in Healthcare, Nutrition, Physical Activity and Psychological Support.

Partners: Fagor Electrodomésticos (Mondragón, Spain), Universidad de Mondragón (Mondragón, Spain), Isoco (Valencia, Spain), Ikerlan (Arrasate-Mondragón, Spain), Hospital Puerta del Hierro (Madrid, Spain), and IBIME-ITACA group of the Universitat Politècnica de Valencia (Valencia, Spain)

FAGOR-DIABETES *Development of a knowledge-based personal health system for the empowerment of outpatients with diabetes mellitus.* Funded by Fagor Electrodomésticos Sdad. Coop. Ltda. (2011-2013)

Objectives: The objective of this project is developing a rule-based CDSS for the healthcare and monitoring of outpatients with diabetes mellitus, aimed to provide a personalized risk assessment based on patient biomedical data and habits related with nutrition and physical activity. The CDSS will gather patient data based on standardized model transformation from the original HIS, and will provide results in a Graphical User Interface showing qualitative and quantitative results associated to the obtained risk assessment.

Partners: Fagor Electrodomésticos (Mondragón, Spain), Universidad de Mondragón (Mondragón, Spain), and IBIME-ITACA group of the Universitat Politècnica de Valencia (Valencia, Spain)

IVI-TWIN-RISK-Prediction *Development of a twin-pregnancy risk predictive model in oocyte donation programme.* Funded by the IVI S.L., and VeraTech for Health S.L. (2009-2012)

Objectives: Aiming to reduce the number of multiple-pregnancy cases in an oocyte donation programme, the objective of this project is to develop a predictive model for the risk assessment of twin pregnancy, providing the probabilities for number of ongoing yolk sacs based on features of the embryo cohort as well as the donor and receipt patients.

Partners: IVI Valencia S.L. (Valencia, Spain), VeraTech for Health S.L. (Valencia, Spain), and IBIME-ITACA group of the Universitat Politèc-nica de Valencia (Valencia, Spain)

The author was involved in additional projects not related with this thesis, including the European project HELP4MOOD, aimed to the intelligent management of patients with major depression; two projects for the development and clinical evaluation of the CDSS CURIAM BT (one National, other funded by the University); a project for the evaluation of the user acceptance of a new HIS in the Balearic Islands; and the National project PRECOG, aimed to a multimedia-based CDSS for ophthalmology.

Additionally, the author performed an international research internship at the Center for Research in Health Technologies and Information Systems (CINTESIS), at the Faculty of Medicine in the University of Porto, Portugal, from 16th September 2013 to 18th December 2013. The research stay was conducted under the supervision of Dr. Pedro Pereira Rodrigues. During the research stay, the author also made research tasks in the Laboratory of Artificial Intelligence and Decision Support (LIAAD), at the INESC-TEC Institute of the University of Porto, under the guidance of Dr. João Gama.

1.5 Thesis outline

The structure of this thesis aims to reflect the different research stages carried out through the development of the thesis. Hence, Chapter 1 has introduced the thesis main motivations, research aims and objectives. Chapter 2 describes the thesis rationale, including the justification of the investigated data quality problems and the required theoretical background. Chapter 3 presents the results of the comparative study of probability distribution distances. Chapters 4 and 5 present the results of the methods for multi-source and temporal variability assessment, respectively. Chapter 6 presents several applications of the developed methods to real case studies in biomedical repositories, with a main focus on the Public Health Mortality Registry of the Region of Valencia. Chapter 7 sets the developed methods in a general biomedical DQ framework, introducing the first steps towards their reuse and industrialization. It is divided in three parts. First, it describes a systematic approach and software for the multi-source and temporal variability methods. Second, it describes the proposal of a general DQ framework including the concepts investigated in this thesis. And third, it compiles three applications that make use of that framework. Finally, Chapter 8 describes the concluding remarks and main recommendations arised from the results of this thesis.

We note that Chapter 6 provides as its introductory notes a simplified summary of the multi-source and temporal variability methods used in the case studies, being supported by the illustrative examples of Appendix D.

Figure 1.1 illustrates an schematic outline of the thesis contributions showing the relationships among the thesis chapters, contributions, publications, projects, stays and the generated software.

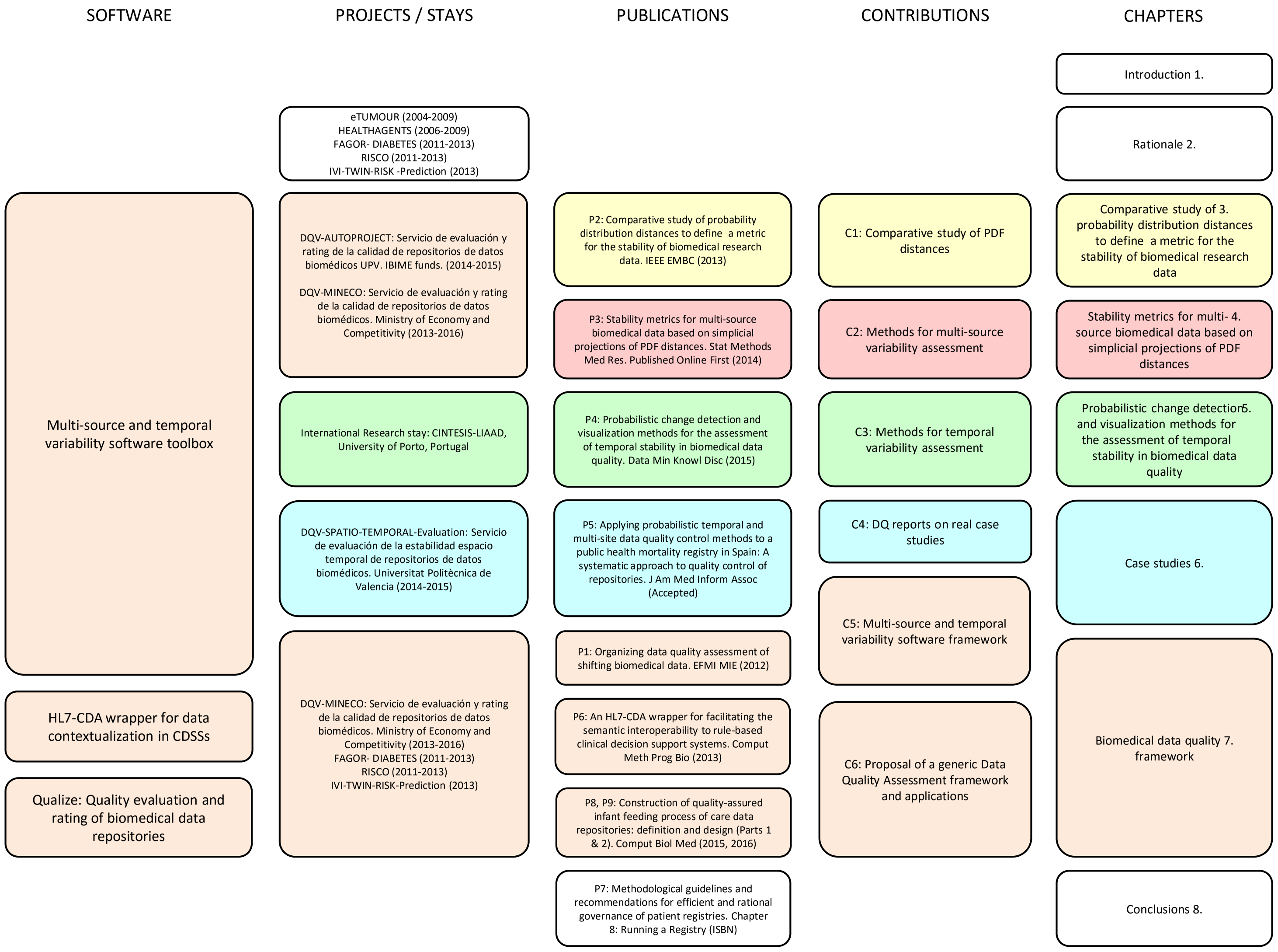

SOFTWARE
PROJECTS / STAYS
PUBLICATIONS
CONTRIBUTIONS
CHAPTERS
Introduction 1.
eTUMOUR (2004-2009) HEALTHAGENTS (2006-2009) FAGOR- DIABETES (2011-2013) RISCO (2011-2013) IVI-TWIN-RISK -Prediction (2013)
Rationale 2.
Multi-source and temporal variability software toolbox
DQV-AUTOPROJECT: Servicio de evaluación y rating de la calidad de repositorios de datos biomédicos UPV. IBIME funds. (2014-2015)
DQV-MINECO: Servicio de evaluación y rating de la calidad de repositorios de datos biomédicos. Ministry of Economy and Competitivity (2013-2016)
P2: Comparative study of probability distribution distances to define a metric for the stability of biomedical research data. IEEE EMBC (2013)
C1: Comparative study of PDF distances
Comparative study of 3. probability distribution distances to define a metric for the stability of biomedical research data
P3: Stability metrics for multi-source biomedical data based on simplicial projections of PDF distances. Stat Methods Med Res. Published Online First (2014)
C2: Methods for multi-source variability assessment
Stability metrics for multi- 4. source biomedical data based on simplicial projections of PDF distances
International Research stay: CINTESIS-LIAAD, University of Porto, Portugal
P4: Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. Data Min Knowl Disc (2015)
C3: Methods for temporal variability assessment
Probabilistic change detection5. and visualization methods for the assessment of temporal stability in biomedical data quality
DQV-SPATIO-TEMPORAL-Evaluation: Servicio de evaluación de la estabilidad espacio temporal de repositorios de datos biomédicos. Universitat Politècnica de Valencia (2014-2015)
P5: Applying probabilistic temporal and multi-site data quality control methods to a public health mortality registry in Spain: A systematic approach to quality control of repositories. J Am Med Inform Assoc (Accepted)
C4: DQ reports on real case studies
Case studies 6.
C5: Multi-source and temporal variability software framework
P1: Organizing data quality assessment of shifting biomedical data. EFMI MIE (2012)
HL7-CDA wrapper for data contextualization in CDSSs
DQV-MINECO: Servicio de evaluación y rating de la calidad de repositorios de datos biomédicos. Ministry of Economy and Competitivity (2013-2016) FAGOR- DIABETES (2011-2013) RISCO (2011-2013) IVI-TWIN-RISK-Prediction (2013)
P6: An HL7-CDA wrapper for facilitating the semantic interoperability to rule-based clinical decision support systems. Comput Meth Prog Bio (2013)
C6: Proposal of a generic Data Quality Assessment framework and applications
Biomedical data quality 7. framework
Qualize: Quality evaluation and rating of biomedical data repositories
P8, P9: Construction of quality-assured infant feeding process of care data repositories: definition and design (Parts 1 & 2). Comput Biol Med (2015, 2016)
P7: Methodological guidelines and recommendations for efficient and rational governance of patient registries. Chapter 8: Running a Registry (ISBN)
Conclusions 8.

Figure 1.1: Outline of the thesis contributions, chapters, publications, projects, stays and software

Chapter 2

Rationale

This chapter describes the thesis rationale divided in two sections. First, the DQ problematic is described, starting from the justification of the discipline, then focusing to biomedical DQ approaches, and ending with the problems of multi-source and temporal variability addressed in this thesis. Second, a general review of the theoretical background recommended for the understanding of the methods developed in this thesis is provided. This review is intended to establish a common basis to complement the descriptions of background and methods explained in the following chapters.

Parts of Section 2.1 were published in the conference paper by Sáez et al (2012b) and the journal publication by Sáez et al (2016)—thesis contributions P1 and P5.

2.1 Biomedical data quality

The outcomes of biomedical research and healthcare practice depend on taking decisions based on the available information (Cruz-Correia et al, 2010). The data behind such information is registered by humans or devices based on observations of facts, at any stage of the healthcare process, and under an environment or context. However, both humans and devices are far from perfect. As a result, errors, omissions, or changes in protocols or practices, can occur during the data acquisition at any of these healthcare process stages or under any context, leading to an unreliable healthcare information caused by a lack of DQ.

Such lack of DQ is an important issue that leads into wrong decisions and suboptimal processes. This is particularly important in the healthcare, where the reliability of information may have direct consequences on the care process of the patients. In primary data use (patient care), low DQ may lead physicians to a set of direct errors, such as inappropriate or outmoded therapy, technical surgical error, inappropriate medication, error in dose or use of medications; and indirect errors, such as failure to take precautions, failure to use indicated tests, avoidable delay in diagnosis, failure to act on results of tests or findings, and inadequate follow up of therapy (Aspden et al, 2004).

Additionally, an insufficient DQ may directly harm the results of studies that reuse the data (Weiskopf and Weng, 2013), such as clinical trials or cohorts. Many of the DQ problems related to the reuse of the clinical information are related to two main causes

(Cruz-Correia et al, 2010): (1) the original Electronic Health Records (EHRs) are designed for its main patient care purpose, without taking into account that further reuse of data that may require different degrees of quality, and (2) those EHRs are not designed envisaging the prevention of DQ problems. Hence, a DQ assessment is important to be aware of such problems for a proper data reuse, improve the value of data and lead to better decisions.

The problem of DQ has been studied for years, specially in the industrial domain, based on the hypothesis that data can be considered a product manufactured by organizations (Madnick et al, 2009)—even though biomedical data in most cases represent a patient's status, data itself is produced by healthcare professionals as well as by devices. Under this assumption, the Massachusetts Institute of Technology (MIT) launched in 1992 the Total Data Quality Management (TDQM) program (Madnick and Wang, 1992), based in the features of Total Quality Management (TQM) introduced in early 1980's for the management of quality in industry. Furthermore, many other research and industrial DQ Assurance proposals have been related to the TQM Six Sigma process improvement methodology (Wang, 1998; Röthlin, 2010; Sebastian-Coleman, 2013). Concretely, the 'DMAIC' model can be used to improve the DQ and their related processes, involving the following cycle of stages: Define, Measure, Analyse, Improve and Control.

DQ Assurance protocols combine activities at different levels, from the design of the information system, the user training in DQ, to a continuous DQ control. Defining what to measure and how to do it is the basis for the DQ Assurance, being them the initial steps to any DQ improvement. There is a general agreement about defining DQ in terms of fitness for purpose (Karr et al, 2006; Madnick et al, 2009), and this can be expressed by the so-called DQ dimensions.

2.1.1 Data quality dimensions

DQ assessment has mostly been defined according to DQ dimensions: attributes that represent a single aspect or construct of DQ (Wang and Strong, 1996). Dimensions can conform to data specifications or to user expectations (Wang and Strong, 1996; Lee et al, 2002; Karr et al, 2006).

The work by Wang and Strong (1996) established a seminal work towards a conceptual framework for DQ assessment considering DQ dimensions. Based on an inductive study from data users opinions, along with a deductive approach based on their experience, they summarized a set of 179 desired data attributes into 15 dimensions, and classified into 4 main groups, and divided into four categories (table 2.1). Since then, many other research studies have aimed to define, compile or suggest, a set of generic DQ dimensions and methodologies for DQ assessment (Wang and Strong, 1996; Lee et al, 2002; Pipino et al, 2002; Pierce, 2004; Oliveira et al, 2005; Karr et al, 2006; Heinrich et al, 2007). We refer to the work by Batini et al (2009) for an extensive review on methodologies for DQ assessment and their relation to dimensions. The main conclusion of most authors is that, in general, there is little agreement about the definition of dimensions and their meaning. However, it can be found that, despite the differences, the proposed dimensions and solutions aim to address conceptually-similar DQ features.

Regarding to the biomedical domain, quality of biomedical data has been studied in routine EHRs (Cruz-Correia et al, 2010; Liaw et al, 2011), in data repositories for study cohorts (Arts et al, 2002; Müller et al, 2003; Bray and Parkin, 2009; Kahn et al, 2012; Walker et al, 2014), and in the integration of heterogeneous sources (Choquet et al, 2010; Cruz-Correia et al, 2010; Kahn et al, 2012).

Weiskopf and Weng (2013) and Liaw et al (2013) reviewed DQ assessment methods and dimensions specially focused to the biomedical domain. Based on an iterative inductive approach, Weiskopf and Weng (2013) performed a systematic review of 95 articles, from which 27 unique terms describing dimensions were obtained. From these, they empirically derived five high-level dimensions: completeness, correctness, concordance, plausibility and currency (Table 2.2). Besides, Liaw et al (2013) used an ontological approach, from which a similar set of five high-level dimensions—among others— was identified as well, including: completeness, accuracy, correctness, consistency and timeliness (Table 2.3). Comparing these two approaches, we observe that both definitions of completeness are compatible. Besides, the definitions of currency and timeliness are compatible among each other as well. Regarding to the other dimensions we observe that, although their definitions seem to be unmatched, in overall the same concepts are covered.

Given that quality is generally associated to fitness for purpose, both Weiskopf and Weng (2013) and Liaw et al (2013) started from little consensus in what and how dimensions are. However, they ended up with a coherent set of high-level dimensions, encouraging further research on discussing these dimensions and establishing proper methodologies for biomedical DQ assessment.

2.1.2 Multi-source and temporal variability

This thesis focuses in the assessment of two problems which, to our opinion, have received insufficient attention as DQ problems, and which the state of the art lacks of appropriate assessment methods: 1) the variability of data distributions among different data sources (e.g., sites or practitioners) and 2) the variability of data distributions through time.

The problem of variability among data sources is generally related to semantic or integration aspects (Knatterud, 2002; Sayer and Goodridge, 2007; Kahn et al, 2012; Walker et al, 2014). However, semantic interoperability does not ensure that variability issues that keep reflected in data probability distributions are properly managed, such as variability in clinical protocols, data acquisition methods or healthcare policies, geographic and demographic differences in populations (Galea et al, 2005), systematic errors, personal or global biases, or even falsified data (Knatterud et al, 1998)—a set of examples in the literature are provided in section 4.2.1).

The multi-source probabilistic variability could be classified as an intrinsic or contextual DQ dimension of multi-site repositories, according to the categories by Wang and Strong (1996), but seems not to fit in any of the dimensions proposed by the same authors. We may fit the multi-source variability to some degree within the descriptions

Table 2.1: Definitions of DQ dimensions by Wang and Strong (1996)

Category	Dimension	Attributes
Intrinsic Data have quality in their own right	Believability	Data are believable
	Accuracy	Data are certified error-free, accurate, correct, flawless, reliable, errors can be easily identified, the integreity of the data, precise
	Objectivity	Unbiased, objective
	Reputation	Reputation of the data source, reputation of the data
Contextual Data quality must be considered within the context of the task at hand	Value-added	Data give you a competitive edge, data add value to your operations
	Relevancy	Applicable, relevant, interesting, usable
	Timeliness	Age of data
	Completeness	Breadth, depth, and scope of information contained in the data
	Appropriate amount of data	The amount of data
Representational Data are presented in an intelligible and clear manner	Interpretability	Data are interpretable
	Ease of understanding	Easily understood, clear, readable
	Representational consistency	Data are continuosly presented in same format, consistently represented, consistently formatted, data are compatible with previous data
	Concise representation	Well-presented, concise, compactly represented, well-organized, aesthetically pleasing, form of presentation, well-formatted, format of the data
Accessibility Data is accessible and secure	Accessibility	Accesible, retrievable, speed of access, available, up-to-date
	Access security	Data cannot be accessed by competitors, data are of a proprietary nature, access to data can be restricted, secure

Table 2.2: Definitions of DQ dimensions by Weiskopf and Weng (2013)—95 articles examined

Dimension	Description
Completeness	Is a truth about a patient present in the EHR?
Correctness	Is an element that is present in the EHR true?
Concordance	Is there agreement between elements in the EHR, or between the EHR and another data source?
Plausibility	Does an element in the EHR makes sense in light of other knowledge about what element is measuring?
Currency	Is an element in the EHR a relevant representation of the patient state at a given point in time?

Table 2.3: Definition of DQ dimensions by Liaw et al (2013)—61 articles examined

Dimension	Descriptions
Completeness	The extent to which information is not missing and is of sufficient breadth and depth for the task at hand. The ability of an information system to represent every meaningful state of the represented real world system. Degree to which information is sufficient to depict every possible state of the task. All values for a variable are recorded. Availability of defined minimum number of records/patient.
Correctness	The free-of-error dimension. Credibility of source and user's level of expertise. Data values, format and types are valid and appropriate; an example is height is in metres and within range for age. Recorded value is in conformity with actual value. Data accuracy includes accuracy and completeness.
Accuracy	Recorded value is in conformity with actual value.
Consistency	Representation of data values is same in all cases. Includes values and physical representation of data. The extent to which information is easy to manipulate and apply to different tasks. The equivalence and process to achieve, equivalence of information stored or used in applications, and systems. The extent of use of a uniform data type and format (e.g. integer, string, date) with a uniform data label (internal consistency) and codes/terms that can be mapped to a reference terminology (external consistency).
Timeliness	Data is not out of data; availability of output is on time. Extent to which information is up to date for task. The delay between a change of the real-world state and the resulting modification of the information system state.

of concordance dimension by Weiskopf and Weng (2013) or within the consistency dimension by Liaw et al (2013). Nevertheless, most specific definitions for consistency aim to whether individual data registries satisfy domain constraints, rules or plausible relations (Cali et al, 2004; Karr et al, 2006).

The aforementioned issues for multi-source probabilistic variability, can appear as well during the time period the data in a repository is being acquired, leading to the second problem: the temporal variability. This is mainly due to the fact that when data is collected for long periods of time, the processes that generate such data, nor the inherent biological and social-behaviour, do not need to be stationary, leading to changes in data distributions. As we will see in Chapter 5, we classify changes in distributions in gradual, abrupt or recurrent, as well as define the concept of temporal subgroups.

The probabilistic temporal variability could as well be classified as an intrinsic or contextual DQ dimension. In this case, time is a factor which has been studied as part of DQ in many works, generally leading to dimensions such as timeliness, currency or volatility. However, these dimensions are generally related to whether individual data registries are up-to-date compared to their real-world values, or what is this rate of change. For a review of time-related DQ dimension definitions we refer to Table I of the work by Heinrich et al (2009) and Table II of the work by Batini et al (2009). Hence, considering a data reuse task in hand, we may consider that the aforementioned changes in data distributions may to some degree affect to up-to-date data is, and therefore relate our temporal variability to the timeliness dimension in the literature. Besides, studies such as clinical trials or public health registries, consider it as concordance or comparability dimensions through time (Svolba and Bauer, 1999; Bray and Parkin, 2009; Kahn et al, 2012).

Multi-source and temporal variability problems, if unmanaged, can result especially harmful in large multi-site reuse repositories, where they can lead to inaccurate or unreproducible results (McMurry et al, 2013; Sáez et al, 2014b, 2015) or even to invalidate them (Kahn et al, 2012). Multi-source variability, as mentioned, is generally related to semantic or integration aspects. However, semantic interoperability does not assure the management of the aforementioned variability problems. Unfortunately, these will keep reflected in data probability distributions.

The reuse of data in multi-site repositories for population studies, clinical trials, or data mining rests on the assumption that the data distributions are to some degree concordant irrespective of the source of data or of the time over which the data have been collected and therefore allows generalizable conclusions to be drawn from the data. Differences in data distributions due to differences in data sources or due to temporal changes, by making the above assumption questionable, may hinder the reuse of repository data and may complicate data analyses, bias the results, or weaken the generalizations based on the data.

Common methods of assessing multi-source variability consist of comparing statistics of populations such as the mean (Bray and Parkin, 2009; Kahn et al, 2012) or comparing the data to a reference dataset (Weiskopf and Weng, 2013). Besides, methods for assessing temporal variability, originally based on quality control of industrial processes, include statistical monitoring used in clinical contexts like Shewart charts (Shewhart and Deming, 1986) or, in laboratory systems, Levey–Jennings charts and Westgard rules (Murray, 1999).

Most of these methods are based on classical statistical approaches, which face two main problems. First, classical statistical tests may not be suitable for multi-type data (e.g., numerical and categorical variables), multivariate data (several variables that change simultaneously), and multi-modal data (distributions generated by more than one component, e.g., data from several disease profiles)—the very characteristics of biomedical data (Sáez et al, 2013b). Second, classical statistical methods may not prove adequate for Big Data (Lin et al, 2013; Nuzzo, 2014; Halsey et al, 2015). Finally, for data from multiple sources, a gold-standard reference dataset may not be available.

As a consequence of the aforementioned problems, it arised the need of investigating proper methods which could simultaneously (1) assess the variability of data distributions among sources and through time while (2) being robust to multi-type, multivariate and multi-modal data, adequate to Big Data and not requiring a reference dataset.

To this end, we first carried out a comparative study to select the proper robust methods to compare probability distributions, which is described in Section 3. Based on the outcomes of such study, we designed and constructed the data quality assessment methods for multi-source and temporal variability, which are described in Sections 4 and 5.

2.2 Theoretical background

With the purpose of improving the assessment of data variability in large, multi-site biomedical data repositories, and considering the aforementioned requirements, in this thesis we have developed two sets of multi-source and temporal DQ assessment methods. They are based on an Information-Theoretic and Geometric framework supported by the measurement of distances among data probability density/mass functions (PDF). Therefore, with the purpose to facilitate the reading and understanding of the following chapters, this section gives an overview to their required theoretical background. The detailed description of these methods is described in the following chapters (Chapters 3, 4 and 5)

2.2.1 Variables and probability distributions

An individual is the unitary entity subject of an study, which belongs to a population of individuals of common features which criteria are defined according to the study. E.g., a patient is an individual of the population of all the possible patients. Individuals are also known as subjects, instances, or cases, among others. A sample is a manageable set of individuals representing the population of study, as it is generally difficult or not possible to account for all the individuals of a population.

Variables and types

A variable—or random variable—is a measurement or observation of an individual's feature, which can take different possible values. To represent such feature in a population, variables are defined as an alphabetic character, which for each individual take their measured or observed value. Typically, variables are denoted by uppercase letters, e.g., X , while their instantiations when their value is not presented, by the corresponding lowercase, e.g., x. As an example, if the Body Mass Index (BMI) of patients in a experiment are represented by X, the specific BMI of patient number i could be $x_i = 22$ (i is as well a variable representing the patient number).

Variables can be of different types. The type of a variable depends first on the nature of the feature, but it can also be set according on how it is measured or observed. Hence, when we can state that we *measure quantitative* variables, and we *observe qualitative* variables. We should note that according to a purpose, an individual's feature could be defined as well as a quantitative or qualitative variable: e.g., given a tumour mass we could measure its size in mm^3 units, or classify this size into an ordinal qualitative category among $\{small, medium, big\}$.

As a consequence, types of variables are mainly divided in those quantitative or numerical and those qualitative or categorical. Numerical variables are mainly divided in discrete and continuous. Numerical discrete are variables which can only be measured in the domain of natural numbers $\mathbb{N}$, namely integers. Numerical continuous are variables which can be measured in the domain of real numbers $\mathbb{R}$, i.e., can have values within two integers. On the other hand, categorical variables are mainly divided in ordinal and non-ordinal. Categorical ordinal are variables on which there is an implicit magnitude value among their possible values. In contrast, in categorical non-ordinal such an order does not exist among their values.

The type of a variable has a great implication in how this variable is analysed in research studies, mainly due to how the frequencies of appearance of their possible values are interpreted at a population level, as described next.

Probability distributions

Probability distributions are mathematical functions which assign a probability of occurrence to the possible values a variable can take. Given a variable X, if x is value and p the variable probability function, then:

$$p(X = x) \to [0, 1], \tag{2.1}$$

where to simplify notation we can assume $p(X = x) \equiv p(x)$.

Such functions, originally depend on the nature of the variable in a population. Hence, several families and specific probability distribution functions exist, which can be parametrized to be adapted to a specific sample. These are known as *parametric distributions*. For a given probability function p, with $\boldsymbol{\Theta}$ as its vector of parameters, then:

$$p(x|\boldsymbol{\Theta}) \to [0, 1]. \tag{2.2}$$

We can classify distributions between continuous or discrete, mainly according to the variable type.

Continuous distributions are those which domain are continuous variables, i.e., their number of possible values is infinite in a range $[a, b]$ (e.g., $[-\infty, +\infty]$), and which derivative through $\mathbb{R}$ is 1, being derivable through all the range (Equation 2.3). The probability function $p(x)$ of continuous variables is known as *probability density function*.

$$\int_a^b p(x)\,\mathrm{d}x = 1. \tag{2.3}$$

In continuous distributions, the Cumulative Density Function (CDF) is a function which given a probability value p obtains the value x, such that p is the probability that a variable X takes a value less or equal than x:

$$CDF(x) = p(X \leq x) = \int_{-\infty}^{x} p(x)\,\mathrm{d}x, \tag{2.4}$$

with $p \in [0, 1]$.

On the other hand, discrete distributions are those which domain are numerical discrete or categorical variables. For numerical discrete, the number of possible values can be countable infinite, e.g., the counts of an event. For categorical variables, the number of possible values is a finite number of elements. In this case, the sum of probabilities of all values sum 1 (Equation 2.5). The probability function of discrete variables is known as *probability mass function*.

$$\sum_{c \in C} p(X = c) = 1 \tag{2.5}$$

The range $[A, B]$ or set C of possible values in continuous and discrete distributions is known as the distribution *support*.

From now, we will use *probability distribution function* (PDF) indistinctly for both probability density functions and probability mass functions. Hence, different PDFs exist for both continuous and discrete distributions, which can be used to better represent specific populations. Probably, the most used PDFs come from the exponential family, which provide a canonical mathematical form based on which several continuous and discrete PDFs can be expressed (Nielsen and Garcia, 2009). Among these we find the Normal and Multinomial distributions, which are the basis for some methods used in this thesis that will be described in this chapter.

The most used exponential family distribution is the Normal distribution, due to its wide representation of real-world successes and to the *central limit theorem*. The Normal distribution is a continuous distribution which support is defined in the range of $[-\infty, +\infty]$, which PDF is ruled by the vector parameter $\boldsymbol{\Theta} = \{\mu, \sigma\}$, and is defined as:

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.6}$$

The parameter μ defines the central tendency or *expected value* of the values of the variable. The probability of observing values below or above such central tendency is symmetric and decay exponentially according to the parameter σ.

Another useful distribution in the exponential family is the Multinomial distribution. The Multinomial distribution is a discrete distribution which models the probability of observing each of a set of k values after a number n of trials. The support of the Multinomial distribution is defined by the number of times each of the k values is observed, and it is ruled by the vector parameter $\mathbf{\Theta} = \{n; p_1, p_2, ..., p_k\}$, where each p_i represents the prior probability of observing the value i in a single trial, with $\sum_i^k = 1$. Hence, the PDF of the Multinomial distribution is defined as:

$$p(x_1, \ldots, x_k | n, p_1, \ldots, p_k) = \begin{cases} \dfrac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, & \text{when } \sum_{i=1}^k x_i = n, \\ \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

The Multinomial distribution can be restricted to other useful distributions by limiting the number of values k or trials n to 1. Hence, $k = 1$ leads to the Binomial distribution, which models the probability of observing a number x of successes in a twice-outcome experiment after n trials. Its PDF is ruled by the vector parameter $\mathbf{\Theta} = \{n, p\}$, where n is the number of trials and p the prior probability of observing one of the two outcomes, generally the positive one, and its PDF is defined as:

$$p(x|n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.8}$$

Besides, limiting the number of trials to $n = 1$ leads to the Categorical distribution. In this situation, Multinomial distribution is sometimes mentioned equivalently to Categorical distribution, i.e., as the probability of observing one success from a set after a single trial. The PDF of the Categorical distribution is ruled by the vector parameter $\mathbf{\Theta} = \{p_1, p_2, ..., p_k\}$, where each p_i represents the prior probability of observing the value i in the trial, with $\sum_i^k = 1$, and its PDF is defined as:

$$p(x_1, \ldots, x_k | p_1, \ldots, p_k) = \prod_{i=1}^{k} p_i^{x_i}, \tag{2.9}$$

where $[x_1, \ldots, x_k]$ is a binary vector where the element at the index of the category to be tested is equal to 1 and the rest to 0.

The last distribution to be introduced is the Beta distribution, a continuous distribution which support is defined in the range of $[0, 1]$, and thus can be used to model measurements ranged within those values (such as proportions or probability values). The Beta distribution is ruled by the vector parameter $\mathbf{\Theta} = \{\alpha, \beta\}$, with $\alpha > 0$ and $\beta > 0$, and related to the shape of the PDF. Hence, the PDF of tbe Beta distribution is defined as:

$$p(x) = \begin{cases} \dfrac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{when } 0 < x < 1, \\ \\ 0, & \text{otherwise,} \end{cases} \tag{2.10}$$

where $B(\alpha,\beta)$ is the Beta function:

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t. \tag{2.11}$$

We have reviewed some parametric PDFs, as functions which map the domain of possible values a variable can take to their probability of occurrence. The vector of parameters $\boldsymbol{\Theta}$ for each PDF is what permits adapting it to specific populations. Given that, in general, it is not possible to measure or observe a whole population, it is therefore not possible to know the true value of the population parameters $\boldsymbol{\Theta}$. As a consequence, we can obtain an estimation $\hat{\boldsymbol{\Theta}}$ of such parameters based on a sample of the population. The most common method for estimating the parameters given a PDF is using its classical Maximum-Likelihood Estimation (MLE) method. Generally speaking, what MLE does is searching the parameters which generate with a higher probability the measured sample, in other words, the parameters which maximize the joint probability of all the sample individuals. The sample joint probability given n individuals of a variable X, is defined as the likelihood function $\mathcal{L}$:

$$\mathcal{L}(X;\boldsymbol{\Theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\Theta}), \tag{2.12}$$

which to avoid numerical computation problems (due to the large product of values near to 0) is generally expressed as the log-likelihood:

$$\ell(X;\boldsymbol{\Theta}) = \log \mathcal{L}(X;\boldsymbol{\Theta}) = \log \prod_{i} p(x_i|\boldsymbol{\Theta}). \tag{2.13}$$

Hence, the MLE can be defined as:

$$\{\hat{\boldsymbol{\Theta}}\} \subseteq \{\underset{\theta\in\Theta}{\arg\max}\ \ell(X;\boldsymbol{\Theta})\}. \tag{2.14}$$

We can observe that there can exist several solutions, depending on the PDF form. Hence, given a specific PDF, the MLE of their vector parameter can be defined as an analytical closed form obtained from the partial derivatives of each parameter or, when this is not straightforward, obtained by means of other optimization methods. As an example, the MLE of the parameters μ and σ of a Normal distribution are the well-known equations for sample mean and sample variance:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.15}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{2.16}$$

with n as the number individuals.

After reviewing several PDF a question arise: what specific PDF should be used for a given variable in a population? It is straightforward that we can first choose between continuous or discrete distributions according to the variable type. Then, for categorical variables the problem of choosing a specific PDF is quite easy, and it comes from the number of categories and trials of what is represented by the variable. However, in general, in real-world experiments with numerical variables it cannot be known a priori which is the inherent generating function of data, i.e., its PDF. Nevertheless, some knowledge about the variable may help choosing one. For many experiments, the Normal distribution is adequate when we expect a central tendency with a degree of variance on the measurements. Besides, a patient survival time can accurately modelled with a Gamma distribution. In such cases we talk about uni-modal variables: variables where there exist a single local maxima, or in other words, a single cluster of measurements around the most repeated value (the statistical mode). This situation is common when we measure a population with a very specific criteria, e.g, in manufacturing industrial processes, where the product specification is the expected value. However, it is uncommon to deal with these types of populations. In fact, in the biomedical sciences the diversity of individuals, such as the patient conditions (e.g., a sample of patients with distinct diagnostics), may entail different inherent generating functions of individuals. In this situation, we talk about multi-modal variables, which can be modelled by means of combining various PDFs, as described next.

Mixture distributions

Mixture distributions are used to model variables which individuals are generated by several inherent functions, including multi-modal variables. Hence, the PDF of a mixture distribution is defined as a weighted sum of several single PDFs, namely mixture components:

$$p(x|\{\boldsymbol{\Theta_1}, \ldots, \boldsymbol{\Theta_c}\}) = \sum_{i=1}^{c} w_i p(x|\boldsymbol{\Theta_i}), \tag{2.17}$$

where $\{\boldsymbol{\Theta_1}, \ldots, \boldsymbol{\Theta_c}\}$ are the vector parameters for each component and w_i the weight for component i, with $\sum_{i=1}^{c} w_i = 1$ in order to maintain the range of the PDF as a probability. When the number of components c is known, we talk about finite mixtures.

Mixture distributions are useful to model variables where a single PDF does not accurately represent the variable value occurrences in a sample. Theoretically, a mixture of Normal distributions may accurately represent most continuous variables. However, in practice the main problem is that there can be too many unknown parameters to estimate: the number of components, their weights, and each of the components' vector parameter. Fixing some of them based on any prior knowledge may simplify the estimation, e.g., when measuring the voxel intensity in brain MR images, the number of mixture components may be associated to the number of brain tissues.

Several methods can be used to estimate the unknown parameters in mixture distributions. Probably the most used is the Expectation Maximization (EM) algorithm (Dempster et al, 1977), which given an initialization of parameters iteratively approaches to a proper solution by means of introducing unobserved latent variables. Besides the known difficulties of selecting a proper initialization, the EM algorithm can be limited in situations when there is not enough knowledge or, in general, not possible to establish the values of some of the mixture parameters. Further, we may also desire to make the fewest assumptions as possible about the underlying generating functions of data. To this end, and relaxing some of these difficulties of mixture distributions, we may choose to use non-parametric distributions.

Non-parametric distributions

Non-parametric distributions are functions which provide PDFs for numerical variables without the need of assuming any parametric PDF model. Hence, they are able to represent multiple modes or *shapes* not possible with other parametric families. Non-parametric distributions can be estimated based on several methods, such as normalized histograms and Kernel Density Estimation (KDE) methods.

Histograms are a widely used exploratory method to visualize how the individuals of a variable are distributed. Concretely, histograms represent the absolute frequencies of individual observations on different non-overlapping and equally-sized intervals through the variable domain, named *bins*. Hence, given m observations of a variable X, its histogram with n bins can be defined by a set of breaking points $\{b_i, \ldots, b_{n+1}\}$ and a set of frequency points $\{f_i, \ldots, f_n\}$ where $f_i = \sum_{j=1}^{m} [b_i \leq x_j < b_{i+1}]$ (using the *Iverson bracket*).

Histograms can act as a non-parametric distribution by using relative frequencies instead of absolute. This can be done by normalizing the counts on each bin by the total number of individuals so that $f_i = \sum_{j=1}^{m} [b_i \leq x_j < b_{i+1}] / m$ and, therefore: $\sum_{i=1}^{n} f_i = 1$. A property of normalized histograms is that they can be interpreted as a Multinomial distribution with a number of trials $n = 1$, where bins—now with an implicit order—represent the possible values observations can take. Hence, this can be used as a method to *discretize* continuous data.

Normalized histograms are a simplistic method which facilitate some operations on distributions, as we will see in further sections. However, its main disadvantage may come from the selection of the proper number of bins to represent data without missing information. That is, too few bins may cause missing information regarding to the variable shape, while too many bins may lead to frequencies composed by very few individuals, being poorly informative or overfitted. To alleviate such problem, several methods exist to calculate the optimum number of bins for a given sample (Silverman, 1986; Guha et al, 2004; Shimazaki and Shinomoto, 2007).

An alternative to histograms are KDE methods, also known as Parzen windows (Parzen, 1962; Bowman and Azzalini, 1997), which can be considered a histogram smoothing method. Given a sample of n individuals, its KDE estimation basically consists in a mixture of n functions called *kernels* normalized by a smoothing parameter h:

$$p(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{2.18}$$

where $K(\cdot)$ is a kernel. Kernels are non-negative, 0-centered, continuous functions which integrate up to 1, hence they can be interpreted as a symmetric PDF centered at 0. In the KDE approach, as shown in Equation 2.18, kernels are centered at each individuals' value. Distinct kernel functions can be used, such as the standard Normal PDF (Equation 2.6) with $\mu = 0$ and $\sigma = 1$. Hence, when using the Normal kernel, the KDE can be considered a finite mixture of Normal distributions where both the number of components, and each component parameters are known, avoiding the aforementioned estimation problems in the EM algorithm. The problem is therefore reduced into choosing the appropriate smoothing parameter h, known as *bandwidth*. Here, the selection of the proper bandwidth is equivalent to the selection of the number of bins in histograms, with similar implications. Nevertheless, similar solutions exist as well as methods for automatically choosing the optimum bandwidth (Silverman, 1986; Shimazaki and Shinomoto, 2010).

Figure 2.1 permits comparing the results of different histograms estimated using the classic histogram and KDE-based methods, with different number of bins and bandwiths.

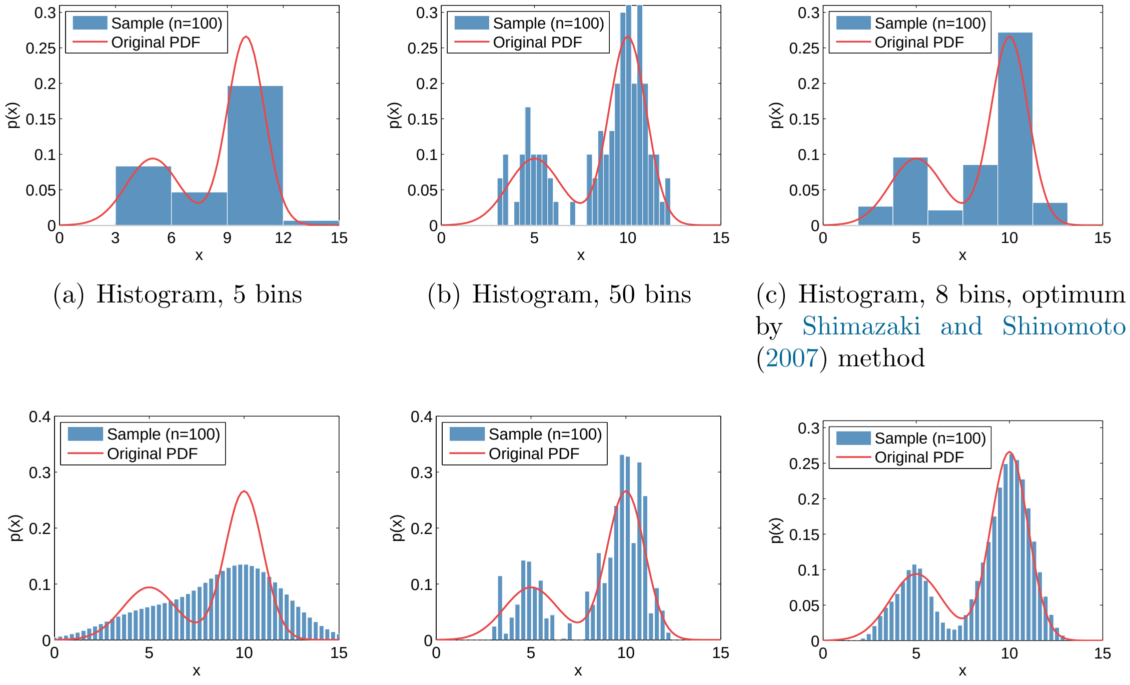
Finally, we should mention that KDE models can be used in practice both as a continuous probability mixture model or as a smoothed normalized histogram. In the latter case, this comes from evaluating the KDE PDF at a set of consecutive equally-spaced points, representing the histogram bins, where the information about the PDF shape will converge as the number of bins increases.

Multivariate distributions

Up to now we have reviewed *univariate* PDFs, i.e., aimed to model a single variable or feature of a population. Besides, when we have several variables in a population, we may model them simultaneously as a *multivariate* PDF—*bivariate* in case of two variables. For d variables $X_1, X_2, \ldots, X_d$, their multivariate PDF can be defined as their *joint distributon*:

$$p(x_1, x_2, \ldots, x_d). \tag{2.19}$$

We first note an important difference between modelling several variables using their joint distribution or using their independent distributions. Two or more variables can be dependent among each other. That is, knowing that an individual has an specific value for one variable, gives us a degree of information about the possible value of other variable. E.g., if we know the height of a patient, we can have some insights about



(a) Histogram, 5 bins

(b) Histogram, 50 bins

(c) Histogram, 8 bins, optimum by Shimazaki and Shinomoto (2007) method

(d) KDE-based histogram, *bandwith* = 2, 50 bins

(e) KDE-based histogram, *bandwith* = 0.1, 50 bins

(f) KDE-based histogram, *bandwith* = .7054, optimum by Matlab© method, 50 bins

Figure 2.1: Results of different histograms estimated using the classic histogram and KDE-based methods, with different number of bins and bandwith. All the histograms have been estimated from the same sample of 100 individuals randomly generated from a bimodal Normal distribution with parameters $\mathbf{\Theta} = \{\mu_1 = 5, \sigma_1 = 2; \mu_2 = 10, \sigma_2 = 1\}$ and weights $\mathbf{w} = \{w_1 = {}^1\!/_3, w_2 = {}^2\!/_3\}$. Estimations with a small number of bins or a large bandwith (a, b) provide a bad representation of the original PDF. Estimations with a large number of bins or a small bandwith provide a noisy histogram, with 0-probability bins. However, automated methods provide better adjusted histograms. We recall that in the KDE case, the shape of the histogram converges as the number of bins increases, as the discrete probabilities are evaluated from a continuous PDF).

her weight. In other words, knowing the height of a patient focus the probability of occurrence of her weight, compared than if nothing was known about the height. Hence, knowing nothing about other variables is equivalent to modelling the variables PDFs independently, while modelling their joint distributions permits modelling the variable inter-dependence. It may happen that the several variables are completely independent among each other, hence, their joint distribution will be equal to their independent distribution:

$$p(x_1, x_2, \ldots, x_d) = p(x_1)p(x_2)\ldots p(x_d). \tag{2.20}$$

Modelling the PDF of joint distributions is not straightforward for several reasons. First, not all families of distributions can be modelled into a single multivariate PDF, specially with mixed types of variables. Second, as the number of variables increase, the domain of the variable and corresponding probabilistic space grows exponentially.

Third, such a large probabilistic space cause that data individuals lie out sparsely, leading to unrepresented variable ranges from which no information can be taken, as part of the known as *curse of dimensionality* problems.

Regarding to the first problem, the multivariate Normal distribution, a widely used multivariate PDF model, can be used as an analytical PDF for several continuous variables. Based on the Normal distribution, it assumes a central tendency for each variable, represented by a vector parameter of means $\boldsymbol{\mu}$, and a covariance matrix Σ. Such a covariance matrix models the independent variance of each variable in its diagonal, while their pairwise covariance in their out-diagonal pairs, the latter representing the aforementioned variable inter-dependence—note that covariances are only pairwise, thus only bivariate dependence is modelled. Hence, given a multivariate vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$, its multivariate Normal PDF is defined as:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)}. \tag{2.21}$$

Knowing that a mixture of Normal distributions, specially in its non-parametric approach based on KDE, is able to accurately represent continuous variables independently of their shape or true family, a KDE-based mixture of multivariate Normals can result useful to model multivariate data problems.

Other simple solution for modelling multivariate data is based on multivariate histograms. One advantage of this solution is that it permits modelling mixed type of distributions, i.e., continuous and categorical data simultaneously. Hence, the histogram domain may contain ordered bins partitioning the continuous space of continuous variables, and bins without an implicit order for categorical data. In any case, this could be modelled as well as a Multinomial distribution with $n = 1$ and k as the total number of bins.

Data sparsity and dimensionality reduction

One of the main problems to the modelling of distributions is that when the number of individuals in the sample is reduced in comparison with the number of dimensions, causing that data points are sparsely distributed in the probabilistic space configured by the dimensions of the problem. This problem is generally known as the *curse of dimensionality*, and it is likely to occur when the probabilistic space is not well represented by the available sample, with higher chances as the number of variables increases, such as in the multivariate solutions mentioned above. Possible consequences of estimating probability distributions in this situation, be parametric or non-parametric, is that the resultant PDF will be with high changes too overfitted to the available sample, causing that new individuals truly generated from the original population may appear anomalous to the estimated model. Similarly, in the case of comparing different low-populated samples of the same population, for example extracted at different moments, when similar estimations would be expected, the measured differences may be very high, leading to biased results.

Data sparsity may impact the results of most data analytics tasks, including those proposed in this thesis. Therefore, data analysts must be aware of this possibility and act in consequence. A general solution to this is reducing the probabilistic space, where a first method is using only the adequate variables for the analysis in hand. Although in some situations the domain of the problem helps in this task (e.g., based on medical evidence), when this information is unknown, automatic feature extraction and selection methods can be used instead (Guyon and Elisseeff, 2003).

Another important solution that can be alleviate to some degree the dimensionality problem is using *dimensionality reduction methods*, which aim to condensate several variables into a manageable transformed lower number of them, maintaining the highest possible of original information:

$$\lim_{m \to n} p(x_1, x_2, \ldots, x_m) = p(x_1, x_2, \ldots, x_n), m < n. \tag{2.22}$$

Hence, the resultant data of a dimensionality reduction method can be modelled as well using any of the PDFs described above.

Dimensionality reduction can be performed based on several approaches, which can be chosen according to the characteristics of the original data. This include linear methods such as Principal Component Analysis (PCA) Pearson (1901), or non-linear such as ISOMAP (Tenenbaum et al, 2000). We should mention that most dimensionality reduction methods are aimed to numerical data where distances among individuals can be computed. Hence, to apply a dimensionality reduction to categorical or mixed variables these can be *encoded* into numerical or binary (Kuhn and Johnson, 2013), or when using embedding-based methods (Cayton, 2005; Lee and Verleysen, 2007), we can establish a distance function among each pair of categories.

Incremental estimation of distributions

The last point of this section is related to the special case when it is not efficient or possible to estimate PDFs using all the sample individuals simultaneously. Suppose that individuals are generated in a timely process, not necessarily at a fixed frequency, and we need to maintain up-to-date a PDF during this time process. With the classical estimation methods, we would need to process all individuals each time the PDF is updated. However, we may not have sufficient computational resources to store all this data in computer memory or make this computation efficiently to provide with the estimated PDF at the time it is required, nor have available enough memory to store all data. Besides, old individuals may not be available anymore, e.g., due to data security/privacy reasons. Similarly, suppose a scenario where a global PDF is to be estimated based on multiple data sources, but external access to their data is restricted. In the aforementioned situations, we may require the use of incremental methods to estimate PDFs.

Incremental estimation methods aim to update the parameters of PDFs by means of adding a new individual or batch of individuals (Jantke, 1993; Cornuéjols, 2010; Gama, 2010; Tortajada et al, 2011). As an example, given a set of samples indexed by i, the parameters of a Normal distribution can be recursively updated by means of only storing three quantities: the past sample size N_{i-1}, the sum of the past observed values $\mathcal{X}_{i-1} = \sum_{j=1}^{N_{i-1}} x_j$, the sum of the squares of the past observed values $\mathcal{X}^2_{i-1} = \sum_{j=1}^{N_{i-1}} x_j^2$:

$$\mu_i = \frac{\mathcal{X}_{i-1} + \mathcal{X}_i}{N_{i-1} + N_i}, \tag{2.23}$$

$$\sigma_i = \sqrt{\frac{\mathcal{X}^2_{i-1} + \mathcal{X}^2_i - \frac{(\mathcal{X}_{i-1}+\mathcal{X}_i)^2}{N_{i-1}+N_i}}{N_{i-1} + N_i - 1}}, \tag{2.24}$$

where N_i is the current batch sample size, $\mathcal{X}_i = \sum_{k=1}^{N_i} x_k$ the sum of the observed values in current batch, and $\mathcal{X}^2_i = \sum_{k=1}^{N_i} x^2_k$ the sum of the squares of the observed values in current batch.

Regarding to the normalized histograms, the incremental estimation of their bin relative frequencies is much more simple. Given a new sample indexed by i, the histogram relative frequencies can be recursively updated by means of only storing the past sample size N_{i-1}. Hence, for a bin j, its relative frequency will be updated as:

$$f_{j_i} = \frac{N_{i-1} f_{j_{i-1}} + \sum_{k=1}^{N_i} [b_j \leq x_k < b_{j+1}]}{N_{i-1} + N_i}, \tag{2.25}$$

where $f_{j_{i-1}}$ is the past bin relative frequency to be updated, N_i is the current batch sample size, and $\sum_{k=1}^{N_i} [b_j \leq x_k < b_{j+1}]$ accounts the absolute frequency of observed values for this bin in the current batch, delimited by the bin breaking points b_j and b_{j+1} (using the Iverson bracket).

Approaches for the incremental estimation of KDE-based distributions have also been investigated (Han et al, 2004; Kim and Scott, 2012; Zhou et al, 2015). Nevertheless, when KDE is used as a smoothing histogram estimation, the aforementioned incremental histogram estimation provide proper solutions as well.

Finally, we note that many other approaches exist to optimize the incremental estimation of PDFs, such as using time windows or forgetting mechanisms. These will be described in more detail in Chapter 5.

2.2.2 Comparing distributions

This thesis investigates methods for the assessment of multi-source and temporal probabilistic variability aimed to the data quality control. Hence, an important aspect is defining what probabilistic variability is and how it is measured.

With probabilistic variability we refer to any dissimilarities among the PDFs of different data sources, or among the PDFs of temporal data batches. Hence, in order to measure, or detect, differences among PDFs we must have the capability to compare such PDFs or more concretely, to measure their differences. While this topic is specially addressed in Chapter 3, where we perform a comparative study of methods for measuring distances among PDFs, in this section we introduce some theoretical background that will help as well throughout the rest of the thesis.

Statistical tests

Probably the most widely used methods by researchers to assess for differences among data samples (or their PDFs) are the corresponding families of statistical tests of hypothesis. Statistical tests of hypothesis are methods aimed to evaluate the evidence about an assumption about one or more populations. Hence, an hypothesis test generally starts from the definition of a *null hypothesis*, H_0, which is to be disproved, being generally the opposite about the assumption to be evaluated. Hence, as Ronald Fisher proposed (Fisher, 1974), assuming that such a null hypothesis is true, the probabilities about getting results at least as extreme as those observed are calculated in the so-called p-value. Consequently, the lower the p-value, the greater possibilities that the null-hypothesis was false.

Several families of statistical tests exist for different purposes. Nevertheless, we are interested in tests for comparing the difference among two or more distributions, which can be separated first in tests for numerical and categorical data, and those for numerical among parametric and non-parametric. Hence, parametric tests are aimed to numerical normally distributed data, while non-parametric tests make no assumptions about the data distribution, and are generally based on individuals ranks or CDF differences. In an attempt to provide a generic description of the theoretical basis for these statistical tests, they basically rely on measuring a *test statistic*, a numerical variable derived from the samples to be compared, which quantifies the proximity to the null hypothesis, and follows a known distribution. As an example, given two samples $X^{(1)}$ and $X^{(2)}$, the Two-Sample Kolmogorov-Smirnov test aims to test whether two non-parametric samples come from the same distribution. Its test statistic is defined as:

$$D_{KS}(X^{(1)}, X^{(2)}) = max|CDF(X^{(1)}) - CDF(X^{(2)})|, \tag{2.26}$$

which measures the maximum difference between the *empirical cumulative distribution functions*—constructed from the continuous increments of the sorted sample individuals—of both samples, where D_{KS} follows the Kolmogorov-Distribution. A measurement of $D = 0$ indicates that the two samples are equal, and thus come from the same distribution. Hence, the test statistic D_{KS} can be considered as a degree of the dissimilarity between the distributions, where the test p-value is generally obtained from reference tables based on D_{KS} and the sample sizes (Conover, 1999).

Even today, the use and interpretation of statistical tests is controversial. Since the beginning of statistical testing, some of its main founders including Ronald Fisher, Harold Jeffreys, Jerzy Neyman and Egon Pearson, disagreed in the procedures and interpretation of statistical tests (Berger, 2003; Nuzzo, 2014). While Fisher introduced the p-values as a measure of evidence, Neyman and Pearson introduced the concepts of statistical power, false positives and negatives, and alternative hypothesis, where a *critical value* as a fixed threshold on a test statistic would lead to the acceptance or rejection of the hypothesis. Besides, Jeffrey advocated for a Bayesian approach for statistical testing.

Some authors affirm that today's common statistical procedures are a hybrid system of those initial frameworks (Goodman, 1999; Nuzzo, 2014). Additionally, an important drawback about the interpretation of those tests, what is specially important in Big Data, is the fact that the larger the sample sizes the easiest is to find statistical differences (Lin et al, 2013), even if the *effect size*—the magnitude of difference in the indicator of study—is not relevant in practice (Sullivan and Feinn, 2012). The miss-use and miss-interpretation of p-values has also been discussed in the healthcare research (Biau et al, 2008; Greenland, 2011; Greenland and Poole, 2013). We will not go into further detail, but leave it here with the purpose to recall the importance of any investigation on methods that could pose an alternative to these classical statistical tests on which these may not be suitable or easily interpretable.

Information-theoretic distances

Suppose two probability distributions $P = p(x)$ and $Q = q(x)$. Information-theoretic distances are functions which measure the distance or dissimilarity between two probability distributions as $D(P||Q)$. These, are mainly derived from Shannon's entropy theory (Shannon, 1948, 2001) and Csizar's f-divergences (Csiszár, 1967; Csiszár, 1972), as we describe next.

Suppose a variable X modelled by a PDF, $p(x)$. If a specific value x_k occurs with $p(x_k) = 1$, we can say that observing such a value gives no information. In contrast, if several values of x occur with $p(x)$ near to 0, their observations are giving us high information. In other words, always observing the same value is not *informative* for an observer, while observing different values it is. According to Shannon (1948), a measure of information from an observation x of $p(x)$ is given by the information function $f(x) = log(^1/_{p(x)}) = -\log p(x)$. Hence, the expected (or mean) information in X is:

$$H(X) = -\mathbb{E} \log p(x) = -\sum_{x \in X} p(x) \log p(x), \tag{2.27}$$

where $H(X)$ is known as the entropy of X, and can be defined as well as the degree of uncertainty about the values the variable can take. Note that entropy is generally defined for discrete variables, although it can be similarly defined for continuous variables as:

$$H(X) = -\mathbb{E} \log p(x) = -\int p(x) \log p(x) \, \mathrm{d}x. \tag{2.28}$$

Hence, the situation of largest entropy would be that where all the possible values take the same probability, while the situation of minimum entropy is that where all the probability is given to a single possible value—with the convention of $0 \log 0 = 0$.

Derived from Shannon entropy, Kullback and Leibler (1951) defined the *relative entropy*, or Kullback-Leibler divergence $KL(P||Q)$, as a measure of information inefficiency of assuming a distribution Q when a true distribution is P (Cover and Thomas, 1991), which is defined as:

$$KL(P||Q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}. \tag{2.29}$$

Equation 2.29 can be seen as a discrete, non-parametric, Kullback-Leibler divergence calculus, which sums through each possible value, or bin, in the common support of distributions P and Q. Besides, the Kullback-Leibler divergence can be calculated analytically for some parametric families of continuous distributions based on analytical forms for d-dimensional Gaussians (Equation 2.30) or approximations for mixtures of Gaussians (Hershey and Olsen, 2007).

$$KL(P||Q) = \frac{\operatorname{tr}\left(\Sigma_Q^{-1}\Sigma_P\right) + \left(\mu_Q - \mu_P\right)^{\top} \Sigma_Q^{-1}(\mu_Q - \mu_P) - d - \log_e\left(\frac{\det(\Sigma_P)}{\det(\Sigma_Q)}\right)}{2\log_e(2)} \tag{2.30}$$

Here, we must note that the Kullback-Leibler divergence is not a true distance, since it is not symmetric nor satisfies the triangle inequality and thus does not accomplish the conditions of a metric:

$$\begin{aligned}
&d(x,y) \geq 0 && \text{(non-negativity)}, && (2.31)\\
&d(x,y) = 0 \text{ if and only if } x = y && \text{(identity)}, && (2.32)\\
&d(x,y) = d(y,x) && \text{(symmetry)}, && (2.33)\\
&d(x,z) \leq d(x,y) + d(y,z) && \text{(triangle inequality)}. && (2.34)
\end{aligned}$$

While this may not suppose any inconvenience according to the purpose, other symmetric information-theoretic distances related to the Kullback-Leibler divergence exist. In fact, the Kullback-Leibler divergence resulted as a special case of the f-divergences, an extension of the aforementioned entropy functional to *relative entropy functionals* (Morimoto, 1963; Ali and Silvey, 1966b; Csiszár, 1967; Csiszár, 1972), which established a canonical form for further distribution divergences (Ullah, 1996; Hero et al, 2001).

Therefore, as a first symmetric alternative to the Kullback-Leibler divergence we can find the Jeffrey divergence (Jeffreys, 1973), as the sum of the two possible directions of the Kullback-Leibler divergence:

$$JF(P||Q) = KL(P||Q) + KL(Q||P). \tag{2.35}$$

The Jeffrey divergence shares with the Kullback-Leibler divergence two properties what may result undesired in some situations. The first is that they are unbounded, and the second is that they are numerically unstable with 0-probability bins–tending to infinity. In this regard, as a bounded, numerically stable true-metric information-theoretic distances we remark the Hellinger (Hazewinkel, 1988) and Jensen-Shannon (Lin, 1991) distances, which are at a small constant among each other (Jayram, 2009).

Concretely, the Jensen-Shannon distance $JSD(P||Q)$, square root of the Jensen-Shannon divergence $JS(P||Q)$ (Endres and Schindelin, 2003; Österreicher and Vajda, 2003), can be directly derived from the Kullback-Leibler divergence as:

$$JSD(P||Q) = JS(P||Q)^{1/2} = \left(\frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M)\right)^{1/2}, \tag{2.36}$$

where $M = \frac{1}{2}(P + Q)$.

We recall that one of the most important practical advantages of using information-theoretic distances for comparing distributions with respect to statistical tests is that the former are distribution independent. They can be used as well for numerical and categorical variables, in uni- or multi-variate settings, and considering the full shape of the variable PDF. Hence they are able to accurately compare multi-modal distributions. These aspects are key for the objectives of this thesis, which accomplishment is evaluated for several distribution comparison methods in Chapter 3.

For further reading on information-theoretic distances we refer to the works by Ali and Silvey (1966a); Ullah (1996); Zhou and Chellappa (2006); Liese and Vajda (2006); Basseville (2010); Cichocki et al (2011). Finally, we must mention other non-probability based method for comparing distributions: the Earth Mover's Distance (EMD), a cost-based method that will be evaluated as well in Chapter 3.

2.2.3 Information geometry

Information geometry is a field which translates the concepts and properties of differential geometry into spaces of probability distributions (Amari and Nagaoka, 2007). Concretely, such spaces of probability distributions are known as *statistical manifolds*, which lie on a Riemannian space.

First we introduce the basic concepts of Riemannian manifolds. A (differentiable) manifold $\mathcal{M}$ can be defined as a space of points which can be connected by a continuous differentiable curve through a coordinate system of D dimensions in $\mathbb{R}^D$, where there exists a one-to-one mapping between a given coordinate in $\mathbb{R}^D$ and a point in $\mathcal{M}$. Riemannian manifolds are those equipped with a—D-by-D—metric tensor g, which facilitates calculating the distance between two points generalizing the Pythagorean theorem to any space. To this end, each point in a Riemannian manifold is bundled with a tangent space, where the inner product between two tangent vectors $< \mathbf{u}, \mathbf{v} >$ is parametrized by g, as $< \mathbf{u}, \mathbf{v} >= \mathbf{u}^T g \mathbf{v}$. Based on this, Riemannian manifolds locally acquire certain properties of affine Euclidean spaces, what permits the global calculus in $\mathcal{M}$ of, e.g., lengths, areas, volumes or angles. Concretely, given a manifold $\mathcal{M}$, we can calculate the shortest distance between two points $\mathbf{p}$ and $\mathbf{q}$—i.e., the *geodesic*—as the minimum curve between those points in the manifold coordinate system $\boldsymbol{\gamma}(t)$ applying the metric tensor $g(\boldsymbol{\gamma})$:

$$D_{\mathcal{M}}(\mathbf{p}, \mathbf{q}) = \min_{\gamma(t)} \int_{t_p}^{t_q} \sqrt{\left(\frac{\partial \boldsymbol{\gamma}(t)}{\partial t}\right)^T g(\gamma) \left(\frac{\partial \boldsymbol{\gamma}(t)}{\partial t}\right)} \mathrm{d}t. \tag{2.37}$$

The metric tensor $g(\gamma)$ will be given as a DxD matrix, where each element g_{ij} establishes the curvature between coordinates i and j. Note that in an Euclidean space, the metric tensor will be the Kronecker Delta δ (a matrix where $\delta_{ij} = 0$ for $i \neq j$ and $\delta_{ij} = 1$ for $i = j$).

In summary, the calculus of a geodesic in $\mathcal{M}$ is guided by local velocity vectors at each point's tangent space, and each infinitesimal distance is given by applying the metric tensor to calculate the inner product between those velocity vectors.

Having described the basic concepts of Riemannian manifolds, we can continue to information geometry. A statistical manifold is a Riemannian manifold which coordinates are the parameters of a given parametric distribution function, and which metric tensor is the Fisher Information Matrix (FIM) of such distribution. This metric tensor is defined as the Fisher Information Metric. The FIM is a DxD positive semi-definite symmetric matrix which, for a specific parametric PDF with vector parameter $\mathbf{\Theta}$ of size D, measure the information that a sample of a variable X contains with respect to each co-parameter Θ_{ij}. As an example, Equation A.6 shows the FIM of the univariate Normal distribution, which being diagonal indicates that the MLE estimates of μ and σ are independent—the derivation of the FIM from a distribution log-likelihood function is described in Appendix A.

$$\begin{array}{cc} & \begin{array}{cc}\mu & \sigma\end{array} \\ \begin{array}{c}\mu \\ \sigma\end{array} & \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{(2\sigma^4)} \end{pmatrix} \end{array} \tag{2.38}$$

Hence, we can translate the concepts of differential geometry to statistical manifolds of probability distributions by using the FIM as the manifold's metric tensor $g(\mathbf{\Theta})$ to calculate the geodesic between two distributions:

$$D_{\mathcal{M}}(P,Q) = \min_{\mathbf{\Theta}(t)} \int_{t_P}^{t_Q} \sqrt{\left(\frac{\partial \mathbf{\Theta}}{\partial t}\right)^T FIM \left(\frac{\partial \mathbf{\Theta}}{\partial t}\right)} \mathrm{d}t, \tag{2.39}$$

where $P = p(x)$ and $Q = q(x)$, instances of PDFs with parameters $\mathbf{\Theta}_P$ and $\mathbf{\Theta}_Q$ (see Figure 2.2). As an example, for the univariate Normal distribution case, an analytical closed form to calculate the FIM-distance based on the distribution parameters solving Equation 2.2.3 is proposed in the work by Costa et al (2005).

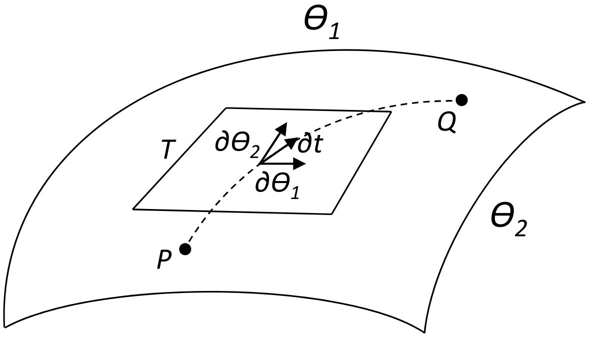


Figure 2.2: Representation of a statistical manifold of a distribution with two parameters Θ_1 and Θ_2 (e.g., $\Theta_1 = \mu$ and $\Theta_2 = \sigma$ for the manifold of a Normal distribution). The geodesic distance between distributions P and Q is calculated differentiating through the curve using the distribution metric tensor guided by the local velocity vectors at tangent space T.

An interesting property of the FIM for information geometry is that it corresponds to the Hessian of the Kullback-Leibler divergence. Hence, the Kullback-Leibler divergence is locally equivalent to the distance based on differentiation through $\mathbf{\Theta}$ using the FIM:

$$KL(\mathbf{\Theta}(\mathbf{t})||\mathbf{\Theta}(\mathbf{t})+\Delta\mathbf{\Theta}) \approx \sqrt{\left(\frac{\partial\mathbf{\Theta}}{\partial t}\right)^T FIM \left(\frac{\partial\mathbf{\Theta}}{\partial t}\right)}. \tag{2.40}$$

Therefore, we can alternatively write Equation as:

$$D_{\mathcal{M}}(P,Q) \approx \min_{\mathbf{\Theta}(t)} \int_{t_P}^{t_Q} KL(\mathbf{\Theta}(t)||\mathbf{\Theta}(t)+\Delta\mathbf{\Theta})\,\mathrm{d}t. \tag{2.41}$$

This property opens the path to the use of the Kullback-Leibler divergence -based f-divergences to approximate FIM-based distances in $\mathcal{M}$, such as the JSD:

$$D_{\mathcal{M}}(P,Q) \approx JSD(P||Q). \tag{2.42}$$

However, given the local equivalence, using these approximations involves a degree of error in larger distances compared to FIM-distances. Nevertheless, this parameter-independent approximation results specially useful when the specific parametrization of the manifold is unknown, as described in the next point.

We note that information geometry has many other applications with parametric families, such as those aimed to model estimation, that will not be discussed for being out of the scope of this thesis. For further reading on Riemannian and information geometry we refer to the and works and books by Petersen (2006); Amari (2001); Csiszár and Shields (2004); Amari and Nagaoka (2007).

Non-parametric information geometry

In many situations the specific family of a set of distributions may be unknown or, as mentioned in previous points, when dealing with multi-variate, multi-modal, and multiple types of variables simultaneously, modelling distributions with specific PDF families may be complicated, in favour of a non-parametric modelling. In these situations, the modelling of a statistical manifold $\mathcal{M}$ with a specific parametrization is not directly applicable. How could we then define such a non-parametric statistical manifold with unknown coordinates? A solution is described next.

A non-parametric statistical manifold $\mathcal{M}$ can be defined based on a set of non-parametric probability distributions $P = \{P_1, \dots, P_n\}$ lied out in $\mathcal{M}$ such as $P_i \in \mathcal{M}$. Then, although the parametrization of $\mathcal{M}$ is unknown, we know that there exists a dissimilarity between any pair of distributions $D_{\mathcal{M}}(P_i, P_j)$. Given that we do not have a FIM, the geodesic distance using the FIM as the metric tensor is not possible. Nevertheless, as introduced before, a good approximation for the geodesic distance is given by the PDF f-divergences. Therefore, based on a specific f-divergence we can measure the $\binom{n}{2}$ pairwise distances among n PDFs in $\mathcal{M}$, which we may represent in a dissimilarity matrix n-by-n Y.

Based on Y we can define a $(n+1)$-dimensional geometrical simplex Δ (see Section 4.3) with $\{P_1, \dots, P_n\}$ as vertices and the corresponding pairwise distances in Y as edges—we define Δ as the maximum-dimensional non-parametric statistical manifold $\mathcal{M}_D$ of dimension $D = n+1$. However, given that Y is based on f-divergences, which are not Euclidean, it is not assured that the simplex edges from Y will fit in the Euclidean space given by $\mathbb{R}^D$. To solve this problem, and with other advantages we mention next, manifold learning algorithms can be used, mainly represented by non-linear Multidimensional Scaling (MDS)-based methods (Torgerson, 1952; Tenenbaum et al, 2000; Cayton, 2005; Borg and Groenen, 2010)—other manifold learning methods include Locally Linear Embedding, ISOMAP (which makes use of MDS), or even PCA as a linear approach. Manifold learning algorithms such as MDS aim to find a D-dimensional Euclidean approximation of a possibly non-Euclidean, unknown-dimensional space of data objects based on the dissimilarities among each pair of objects—see next section for further information about MDS. As a consequence, MDS not only permits estimating our statistical manifold at its simplicial maximum dimensionality, but also can estimate proper 2-dimensional or 3-dimensional projections of $\mathcal{M}$ which permit both its visualization and further efficient calculus in lower dimensions. Further, even when the specific parametric family of PDFs is known, we can apply MDS to translate $\mathcal{M}_\Theta$ into an Euclidean space or obtain a visualization.

As an example, Figure 2.3 shows the non-parametric statistical manifolds formed by a set of Normal PDFs, where pairwise distances have been measured using either the analytical closed form for Normal distributions by Costa et al (2005) and the Jensen-Shannon distance. The 2D and 3D projections have been obtained using MDS from the measured pairwise distances.

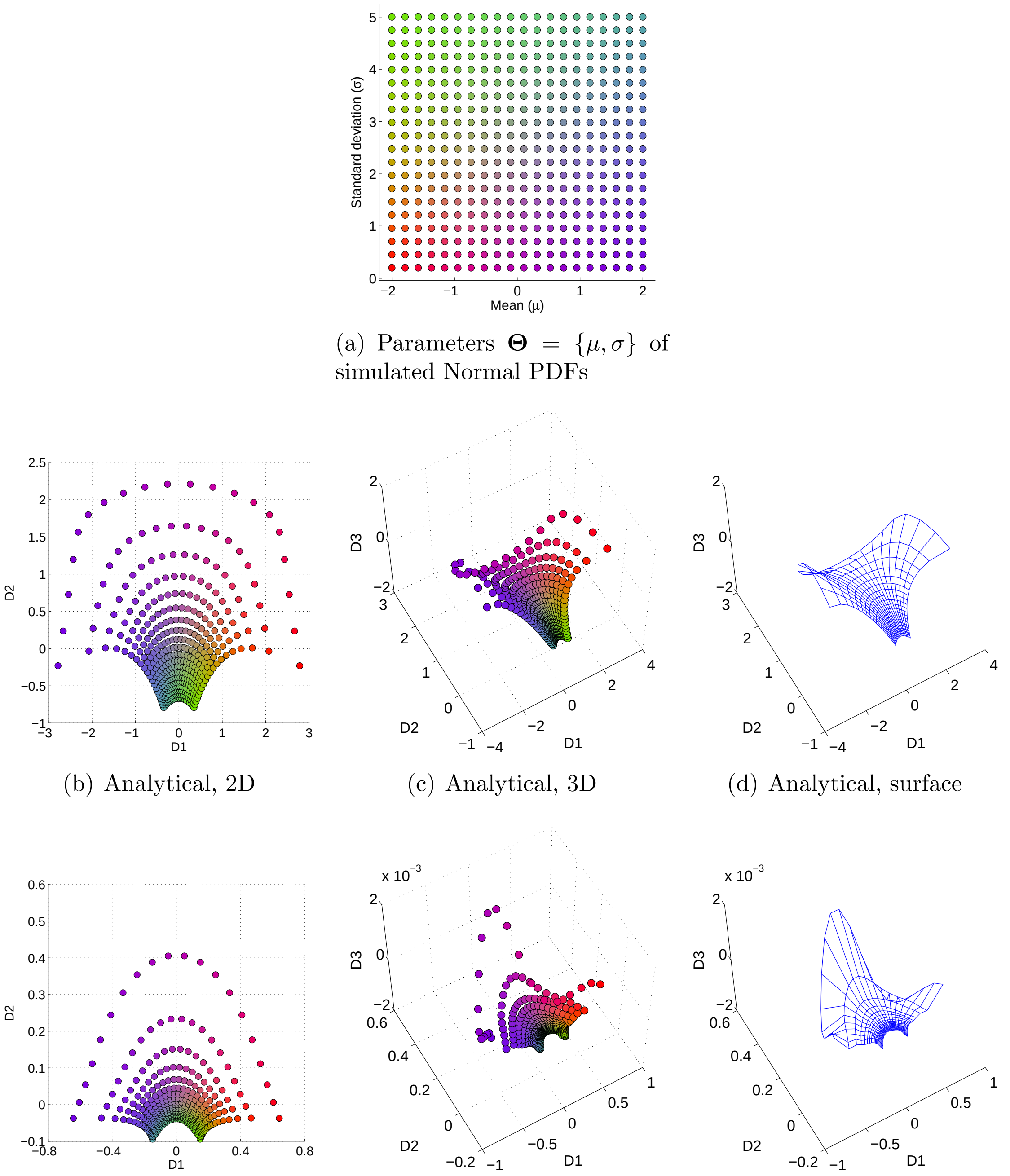
Having an Euclidean representation of the points representing distributions facilitates treating distributions as individuals where the features measured by their dimensions conserve a great degree of information about the full distribution shape. This facilitates in a great measure performing data analytics tasks such as classification or clustering of data distributions.

Further discussion on non-parametric information geometry is shown in the following section as well as in Chapters 4 and 5, where this methodology was applied as the probabilistic framework for the multi-source and temporal variability assessment methods. Concretely, in Chapter 4 an statistical manifold is obtained from the distributions of multiple data sources to build multi-source variability metrics. In Chapter 5 we introduce temporal dynamics in the statistical manifold, what, to our knowledge is the first attempt doing so. For further reading on non-parametric information geometry we refer to the works by Carter et al (2008) and Sun and Marchand-Maillet (2014).

2.2.4 Multi-dimensional scaling

Given a dissimilarity matrix $Y = (y_{11}, \ldots, y_{nn})$ among n points, the objective of MDS is to obtain the set $P = (\mathbf{p}_{11}, ..., \mathbf{p}_{nc})$ of points in a $\mathbb{R}^c$ Euclidean space such that $c \leq n-1$. This is done by finding the best approximation of $\|\mathbf{p}_i - \mathbf{p}_j\| \approx f(y_{ij})$, where $\|\cdot\|$ is the euclidean norm between points $\mathbf{p}_i$ and $\mathbf{p}_j$, and $f(y_{ij})$ is a transformation of the original dissimilarities (optimally $f(y_{ij}) = y_{ij}$). This approximation can be solved by the minimization of the raw loss function:

$$\min_P \sum_{i<j} (f(y_{ij}) - \|\mathbf{p}_i - \mathbf{p}_j\|)^2, \tag{2.43}$$



(a) Parameters $\mathbf{\Theta} = \{\mu, \sigma\}$ of simulated Normal PDFs

(b) Analytical, 2D

(c) Analytical, 3D

(d) Analytical, surface

(e) JSD, 2D

(f) JSD, 3D

(g) JSD, surface

Figure 2.3: Visualization of the non-parametric statistical manifolds formed by a set of 400 univariate Normal PDFs. Pairwise distances have been measured using both the analytical closed form for Normal distributions by Costa et al (2005) and the Jensen-Shannon distance (JSD). The 2D and 3D projections were obtained using MDS from the pairwise distances. The spheres in the plots for statistical manifolds are coloured according to the color of their distributions in (a). In this manner, it can be noted the non-linear relationship between the scalar changes in the parameters and the distances between the corresponding distributions. Concretely, although with different mean and fixed standard deviation neighbour distributions are approximately equidistant, fixing the mean and changing the standard deviation makes that the larger the standard deviations the smaller the dissimilarity among distributions (more probability density is shared). Note that the closed-form is unbounded while the JSD is bounded between 0 and 1. Inspired by the work by Cranmer (2014).

where generally a calibrated loss function such as the Kruskal's Stress-1 (Kruskal, 1964) is used:

$$\min_P \sqrt{\frac{\sum_{ij}(f(y_{ij}) - \|\mathbf{p}_i - \mathbf{p}_j\|)^2}{\sum_{ij}(\|\mathbf{p}_i - \mathbf{p}_j\|)^2}}. \tag{2.44}$$

Modern MDS methods can be classified into metric and non-metric (Borg and Groenen, 2010). In metric MDS the resultant inter-point distances are related to the input dissimilarities by $f(y_{ij})$ as a continuous function, while in non-metric the objective is to preserve the rank order among the dissimilarities with $f(y_{ij})$ as a monotonic transformation. Both methods compute iteratively the best approximation minimizing the Stress functions, starting from an initial configuration of points. If such initialization is obtained using *classical scaling*, based on eigendecompositions, the resultant coordinates will likely be ordered monotonically by their significance with respect to the approximation.

It is important to recall that the smaller the output dimensionality c is chosen, the larger the transformation error, or stress, will be obtained. The special case of $c = n - 1$ is known as full-dimensional scaling, and it provides a perfect embedding when the input dissimilarities are Euclidean, leading to zero Stress. In addition, the solution of full-dimensional scaling is a c-simplex, which can be found in a unique global minima (De Leeuw, 1993).

As introduced in section 2.2.3, MDS is suitable for embedding the non-Euclidean spaces of statistical manifolds into Euclidean ones. In information geometry we can use the FIM-based distances—when known—or their integrated versions (e.g., the JSD) as non-Euclidean metrics between probability distributions. As a consequence, given the change of space and dimensionality by MDS the relationship between the original metric and the embedded one is twofold. First, the change of space may be non-isometric, i.e., it may imply a loss of precision, leading to a positive Stress. And second, the change of space may be isometric, i.e., the original distances are preserved, leading to a zero Stress. Concretely, the Whitney embedding theorem (Whitney, 1940) states that any M-dimensional smooth manifold can be smoothly embedded into an $\mathbb{R}^{2M}$ Euclidean space preserving the neighbourhood, however it does not imply isometry. Besides, Nash embedding theorem(Nash, 1956) guarantees that any Riemannian manifold can be isometrically embedded into an Euclidean space, with some restrictions on the number of dimensions. Hence, whether we get an isometric embedding or not will depend on the number of dimensions c to which we transform, on the original dimensionality of the statistical manifold, and on the number of points being mapped.

For example, the case of non-parametric information geometry, the original dimensionality of the statistical manifold is unknown.[a] Therefore, a sufficiently large number of dimensions will probably lead to an isometric embedding. This may be the case for the full-dimensional scaling mentioned above, when the number of points is as well large. Besides, when the number of points being mapped is low, the topological restrictions (e.g., neighbourhood) are less, therefore, the full-dimensional scaling may

[a] The parametrization of the categorical or mixture distributions could be studied when using a non-parametric estimation of distributions (Section 2.2.1).

provide as well an isometric mapping (e.g., in all the embeddings carried out in Section 4.5, a zero Stress was obtained), or at least provide the closest possible embedding to an isometric one.

Chapter 3

Comparative study of probability distribution distances to define a metric for the variability of multi-source biomedical data repositories

Biomedical data repositories are often composed by data from multiple sources and acquired during long periods of time. In some cases, data may present dissimilarities among their probability density functions (PDF) due to different variability causes among data sources or over time. This may hinder the data reuse when treating data as a whole. Additionally, the overall quality of the data is diminished. With the purpose of developing a generic and comparable metric to assess the variability of datasets among sources and over time, this chapter studies the applicability and behaviour of several PDF distances over shifts on different conditions (such as uni- and multivariate, different types of variable, and multi-modality) which may appear in real biomedical data. From the studied distances, we found information-theoretic based to be the most practical distances for most conditions. We discuss the properties and usefulness of each distance according to the possible requirements of a general variability metric.

The contents of this chapter were published in the conference paper by Sáez et al (2013b)—thesis contribution P2.

3.1 Introduction

Research biobanks are often composed by data from multiple sources (different hospitals, health services, physicians, etc) or acquired during a period of time. A common research task consists in developing a hypothesis or model based in the whole set of multi-source data. However, dissimilarities in the probability density function (PDF) among the different subsets of data or over time may complicate such research, lead to wrong hypothesis, or harm the further use of results on new data. In addition, detecting such dissimilarities may be difficult due to the heterogeneous conditions present in biomedical research data: (1) variables of different types (categorical, ordinal or not; and numerical, continuous or discrete), (2) data coming from uni-modal or multi-modal distributions, and (3) univariate or multivariate data. We classify the presence of such dissimilarities in PDFs as data quality problem related to the multi-source variability of data.

Providing accurate information about the data variability may help data managers and researchers to take decisions during the definition and development of research studies, as well as to feedback data providers about their acquisition procedures. In addition, a generic metric comparable among different studies, may provide a measurement of the degree of variability of biomedical data repositories as a DQ metric.

In this chapter, we study the applicability and behaviour of several pairwise PDF distances on a set of simulations of data shifts based on the aforementioned biomedical data conditions. These pairwise distances provide information about the variability between pairs of sources or time batches. Hence, this study is the first stage towards the development of a global variability metric for any arbitrary number of sources or time batches, where pairwise PDF distances will serve as baseline measurements. We present the results of such comparative study as well as a discussion aimed to the next research steps.

3.2 Background

Weiskopf and Weng (2013) reviewed several studies on biomedical DQ. Most of them focused on measuring DQ dimensions of a data repository as a whole. The concept of data source agreement was found aligned with the problem of multi-source variability we are focusing. Besides, dataset shifts have also been related to DQ problems (Cruz-Correia et al, 2010; Sáez et al, 2012b). Dataset shifts are dissimilarities in the underlying distributions of data which can be originated through the course of time or across multi-source factors. Our aim is to assign a distance to dataset shifts among several sources of data and over time, as a measurement of the overall agreement in data inherent concepts.

Most studies aim to detect dataset shifts in data streams, e.g. based on specific statistical tests (Kifer et al, 2004) or distributional divergences (Dasu et al, 2009). Some of these approaches can be suited to obtain dissimilarity measures among the PDF of different data sources. Some works have also been published comparing PDF dissimilarity measures (Liu et al, 2008; Budka et al, 2011), although aimed to image retrieval. To the best of our knowledge, no similar comparisons have been carried out to assess the variability among biomedical data distributions, envisaging the multi-source, multivariate, multimodal and multi-type conditions, as well as the adequateness to a global variability metric.

3.3 Methods

3.3.1 Simulation

We evaluated the distances on a set of simulations to cover: (1) variable types, (2) multi-modality, and (3) dimensionality. We focused on numerical and categorical data, the most common post-processed research data, which facilitate the statistical analysis. In each simulation, two random datasets, (a) and (b), were defined following the same statistical distribution, where a null dissimilarity is expected. Then, we sequentially increased their dissimilarity until a predefined maximum state, where a maximum dissimilarity is expected. Distances were measured at each dissimilarity level.

We started evaluating the effect of shifts in different univariate variable types, covering (1) and (2). Simulation U1 consisted in a Normal $N(\mu, 1)$ continuous variable (cont.v.) where dataset means $\mu^{(a)}$ and $\mu^{(b)}$ separated each other—e.g. due to an acquisition device that becomes biased. U2: $N(\mu, 1)$ cont.v. where dataset b becomes bi-modal as a mixture of two Normal PDFs defined as $\frac{1}{2}\sum_{c=1,2} N(\mu_c^{(b)}, 1)$, which component means $\mu_1^{(b)}$ and $\mu_2^{(b)}$ symmetrically separate from the original—e.g. due to the appearance of a new pathological pattern. U3: Chi-squared $\chi^2(k)$ cont.v. where degrees of freedom $k^{(b)}$ separated from $k^{(a)} = 0$—e.g. due to an increase in the occurrence of a biomarker. U4: Binomial $B(1, p)$ ordinal categorical variable (cat.v.) where $p^{(a)} = p^{(b)} = 0.5$ shifted to 0 and 1 respectively—e.g. due to variation in gender percentages in a diagnostic group. U5: Multinomial $Mult_3(1, \boldsymbol{p})$ non-ordinal cat.v. which priors shifted from an equal to a maximum difference state—e.g. due to a variation in the number of uses of treatments.

The multivariate simulations consisted in a combination of the previous variables, completing then (1), (2), and (3). M1: bivariate $N(\boldsymbol{\mu}, 1)$ cont.v. which means separated respectively. M2: bivariate $N(\boldsymbol{\mu}1)$ cont.v. where dataset b becomes multimodal as a mixture which component means symmetrically separated from the original. M3: two $B(1, p)$ cat.v. where $p_1^{(a)} = p_2^{(a)} = p_1^{(b)} = p_2^{(b)} = 0.5$ shifted to $p_1^{(a)} = p_2^{(a)} = 1, p_1^{(b)} = p_2^{(b)} = 0$. M4: a combination of a $N(\mu, 1)$ cont.v. with a $B(1, p)$ cat.v. combining the shifts of U1 and U4.

Figures 3.1 and 3.2 shows the initial and final states of the simulated distributions for the univariate and multivariate experiments, respectively.

3.3.2 Estimation of probability densities

To ensure the applicability to any non-parametric continuous PDF, we estimated empirical PDF histograms of the compared datasets using a Kernel-Density Estimation smoothing method (Parzen, 1962; Bowman and Azzalini, 1997), with Gaussian kernels and establishing the optimum bandwith based in the method by Shimazaki and Shinomoto (2010). Additionally, to homogenize the support, we estimated the common PDF from both datasets, and then, its bin centers were used as reference to interpolate the PDF of the independent datasets.

3.3.3 Studied distances

PDF distances measure how far two statistical distributions are in a metric space. As shown in Equation 2.31, a distance metric must be (i) non-negative, (ii) zero only if the two compared distributions are the same (identity), (iii) symmetric, and (iv) must satisfy the triangle inequality. Divergences also provide a measure of dissimilarity, however, do not require to be symmetric nor satisfy the triangle inequality. Distances are then consistent with our purpose of a generic and comparable variability metric.

One type of studied distances included the statistics obtained in classical two-sample statistical hypothesis tests, including the parametric Student's t from t-test and the non-parametric Kolmogorov-Smirnov test statistic, Kruskal-Wallis difference in mean ranks and the obtained χ^2 statistic from the Kruskal-Wallis test. We discarded the χ^2 test statistic for categorical data because it does not accomplish the identity of indiscernibles condition of a metric. Despite these type of distances are conceived for univariate[b] numerical data, we kept these tests for two reasons. First, we want to compare their behaviour in univariate multi-modal data. And second, dimensionality reduction of multivariate datasets may lead to a univariate sample making these methods feasible. Other advantage is that these statistics can be directly associated to p-values which permit significance tests on the differences.

We also studied information-theoretic based distances, which derive from Shanon's entropy theory (Cover and Thomas, 1991), including the Jeffrey divergence and the square root of the Jensen-Shannon divergence, both symmetrized versions of the Kullback-Leibler divergence, the second also smoothed. We also studied the Hellinger distance, which can be defined as a metric version of the Bhattacharyya distance, commonly used in Pattern Recognition. These distances belong to the family of f-divergences (Ali and Silvey, 1966a; Csiszár, 1967), which measure the difference between PDFs. The main advantage of these metrics to the aforementioned statistics is that they apply to any type of binned PDF. Jeffrey and Jensen-Shannon, nevertheless, cannot be measured when any of the PDFs has 0-probability bins—e.g. a categorical value not present in a source—, hence, in such cases an absolute discounting method (Ney et al, 1994) was used to smooth the estimated PDFs.

Finally, we studied the Earth Mover's Distance metric (EMD, equivalent to the Mallows or Wasserstein distance) (Rubner et al, 2000). EMD calculates the minimum cost required to transform one PDF into the other, using a predefined cost matrix of the probability mass flow between the bins in the support (ground distances). Originally conceived for image retrieval, EMD has also been used to measure dissimilarities in multidimensional distributions (Applegate et al, 2011; Dasu and Loh, 2012). EMD envisages inter-bin information, in contrast to information-theoretic distances which make bin-by-bin comparisons, however, involves a higher computational cost. Additionally, EMD relaxes possible losses of information caused by binning, and permits defining custom cost matrices. To adapt the multivariate experiments to EMD algorithm, we embedded the two dimensions into one histogram using a normalized *L1* ground distance matrix.

[b] Bivariate Kolmogorov-Smirnov test approaches (Lopes et al, 2007) require further study since in d-dimensions imply $2^d - 1$ possible orderings. MANOVA tests entail a linear combination of the two or more normally-distributed dependent variables.

3.4 Results

Figure 3.3 shows the results of the experiments. Each distance was normalized between zero and one to facilitate the comparison. As expected, in all simulations the evaluated distances behave monotonically increasing. In addition, all distances begin in 0, and we can observe that while most converge in continuous tests, these are approximately linear in discrete.

In experiment U1 non-parametric statistics behave similarly, converging around a distance between means of 4σ. Information-theoretic distances behave similarly, except Jeffrey divergence, which begins convex and converges when the tails of the PDFs leave each other. The t-test statistic behaves linearly, since we are separating two Normal PDFs with equal variance. The EMD resultant function also converges, but later.

In U2, t-test and Kruskal-wallis statistics were not able to capture the bi-modal shifting—despite the dissimilarity, sample means were the same—resulting in zero. The rest of distances behave equivalently to U1, capturing the bi-modality.

In U3, distances behave similarly to U1, but as it can be appreciated in t-test series, PDF means did not vary linearly with the shift in degrees of freedom.

Categorical simulations U4 and U5 resulted in equivalent results in information-theoretic and EMD distances. However, statistics distances were not applicable to U5, since non-ordinal categorical. Thus, we only show the results of U4, where the first are equivalent. Due to the linear shift in probability masses in these categorical experiments—in contrast to when separating Normal PDFs—none of the distances resulted convex. In addition, some captured this linear density shift resulting in linear functions. Despite the smoothing, we can observe in the Jeffrey distance series the tendency to infinite with smoothed 0-probability elements in the last iteration.

Results of multivariate experiments M1 and M2 are equivalent to their univariate relative U1 and U2, with the exception that statistic tests were not applicable. Thus, all distances converge, although EMD does later. Analogously, the results in Binomial experiment M3 are equivalent to those in U4. We can appreciate, however, slight differences in the results of mixed variable types experiment M4: while Jensen-Shannon and Hellinger distances seem to average the results of its independent continuous and categorical shifts, the EMD transformation cost seem to be slightly higher across the central iterations due to the abrupt density flow through the categorical dimension—we recall that EMD envisages inter-bin information.

3.5 Discussion

We come back to the studied conditions: (1) variable types, (2) multi-modality, and (3) dimensionality. Biomedical data can be considered heterogeneous and multi-modal by nature. Even univariate data may be formed by different 'natural' components, such as a mixture of healthy and different components of unhealthy parameters, or 'artificial' components, such as differences in the quality of data among their generating sources. Thus, an effective distance must be able to capture the dissimilarity in any of these conditions.

Regarding to the evaluation of (1), only information-theoretic and EMD are suited to any type of variable—statistics are only to numerical. Additionally, EMD is the only distance which permits setting specific costs to the difference between categories in unordered categorical data. Regarding to (2), t-test and Kruskal-Wallis had problems detecting multi-modality (U2, M2), however, Kolmogorov-Smirnov, information-theoretic and EMD were successful. Thus, despite the advantage of information-theoretic and EMD in (1) (and as we will see next, in (3)), Kolmogorov-Smirnov might still be used for obtaining a p-value on the difference in a continuous univariate variable resultant from a possible dimensionality reduction on multi-type data. At this point, information-theoretic and EMD distances seem the most practical for most situations. From these, we may consider the issue with null-probability elements of Jeffrey divergence a reason for discarding it. We can also observe that Jensen-Shannon and Hellinger are within a small constant each other (Jayram, 2009). Additionally, as we already mentioned, EMD is able to capture inter-bin information, and it is possible to define any cost between them, what may be useful in categorical data or when grouping PDF signatures (Rubner et al, 2000).

Finally, regarding to the evaluation of (3), we already mentioned that statistics distances were not suited to multivariate data. In contrast, information theoretic distances and EMD are theoretically suited to any number of dimensions. However, direct estimation of PDFs in high-dimensional biobanks may be impractical due both to computational requirements and sparsity in the probabilistic space. Hence, dimensionality reduction methods may be applied to make feasible low-dimensional distances. For instance, we could reduce the dataset into a lower-dimensional manifold. Additionally, in massive-data environments, we could represent groups of similar cases based in PDF signatures to facilitate the distance calculus.

On the other hand, results show that, in general, most distances have a convergence limit. They converge when the volume of the joint density between the two PDFs is minimized converging as well. However, EMD does later, what may suppose two advantages. First, it behaves approximately linear until the saturation level of those that converge first. And second, it can still express dissimilarity farther from this level. Furthermore, a bounded PDF support, e.g. in categorical data or bounded continuous, obviously entails a maximum limit in all the distances. Under these assumptions, we may choose between using the Jensen-Shannon, Hellinger or EMD, depending on the dissimilarity level at which we need the distance to converge. Table 3.1 summarizes the findings.

To be generic, pairwise measurements should provide a dissimilarity level comparable across different datasets, or even different domains—imagine we wish to provide a variability mark in a DQ consulting. Jensen-Shannon, Hellinger, and Kolmogorov-Smirnov distances are bounded by definition between zero and one, what applies here (Kolmogorov-Smirnov, however, did not achieve its maximum value in the bimodal experiment (U2)). On the other hand, we noticed that the normalization applied to the EMD ground distance matrix, where a maximum cost of 1 is given when moving density between extreme bins, makes comparable the resultant transformation cost. This solution, however, requires predefining the possible support of all variables in order to identify the maximum inter-bin costs—equivalent to establishing the bounds of the probabilistic space.

Feature	**T**	**KW**	**KS**	**JF**	**JSD**	**EMD**
Multivariate	-	-	-	Yes	Yes	Yes
Multi-Type	-	-	-	Yes	Yes	Yes
Multi-Modal	-	-	Yes	Yes	Yes	Yes
Bounded	No	No	Yes	No	Yes	Yes

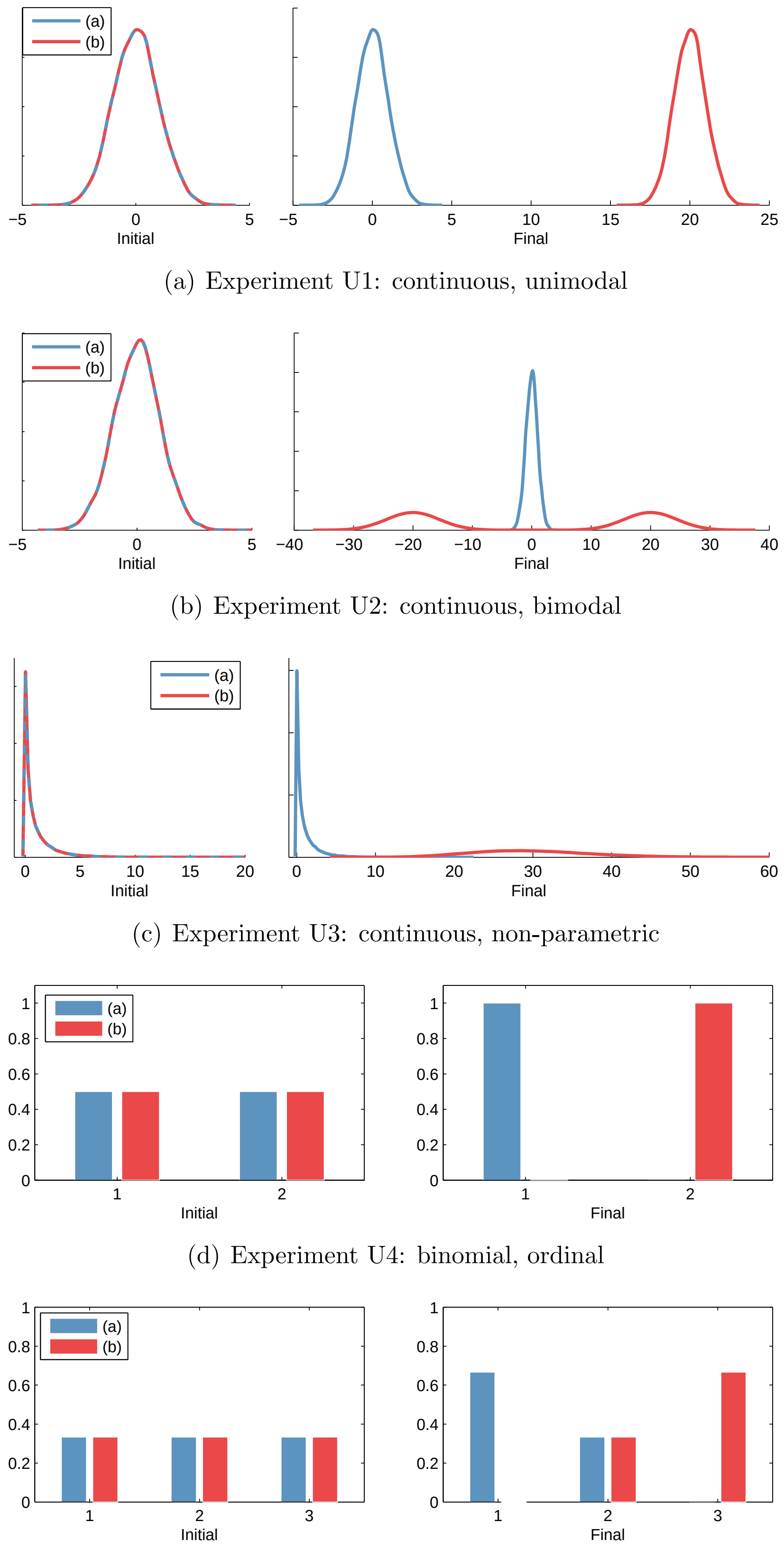
Table 3.1: Ability of PDF distances or test statistics (columns) for dealing with specific features of data (rows 1-3) and whether the distance is bounded (row 4). T: *t*-test statistic, KW: Kruskal-Wallis statistic, KS: Kolmogorov-Smirnov statistic, JF: Jeffrey (Symmetric Kullback-Leibler divergence), JSD: Jensen-Shannon$^{-1/2}$, EMD: Earth Mover's Distance. The '-' means that the corresponding distance is not designed for the corresponding feature.

We have not focused on other common types of biomedical data such as free text, signals or images. In some research tasks, a specific preprocessing may be used to obtain quantitative or qualitative measurements which will permit the use of the methods presented in this work. For instance, the Quantitative Magnetic Resonance (MR) methodology (Wagnerova et al, 2012) is based on different quantitative parameters from brain MR images or MR spectroscopy signals, which may be used to assess the variability among radiology data sources, or among segmented brain tissues.

3.6 Conclusions

Providing information about the variability of biomedical research data among its sources and over time may be of crucial importance. We have studied the behaviour and application of pairwise PDF distances on simulations of multi-type, multi-modal and multivariate conditions of biomedical data. The evaluated distances based on classical statistical tests are only suited to numerical univariate data, and have difficulties in multi-modality. Information-theoretic distances and EMD can handle multivariate, both continuous and discrete, and mixed types data. In general, all distances converge when the joint probability mass between the compared PDFs converges to the minimum.

Therefore, from the studied distances, only Jensen-Shannon and EMD satisfy both the applicability to conditions of biomedical data and the bounding. The EMD permits setting custom inter-bin costs, what allows bounding the metric, however, it requires knowing a priori the bounds of the probability support of all the involved variables. In contrast, the JSD bounds are approached based on the degree of overlapping between the compared distributions (as the overlap decreases the distance tends to its upper bound), defining the distance only by what is measured, avoiding external configurations. As a consequence, and although both EMD and JSD could be suitable for the purpose of a variability metric, the JSD was chosen for its direct foundations in information-theory, what permits constructing over the theory of a probabilistic framework, and for its generalization for comparability. These results establish the basis for the next chapters towards multi-source and temporal variability metrics.



(a) Experiment U1: continuous, unimodal

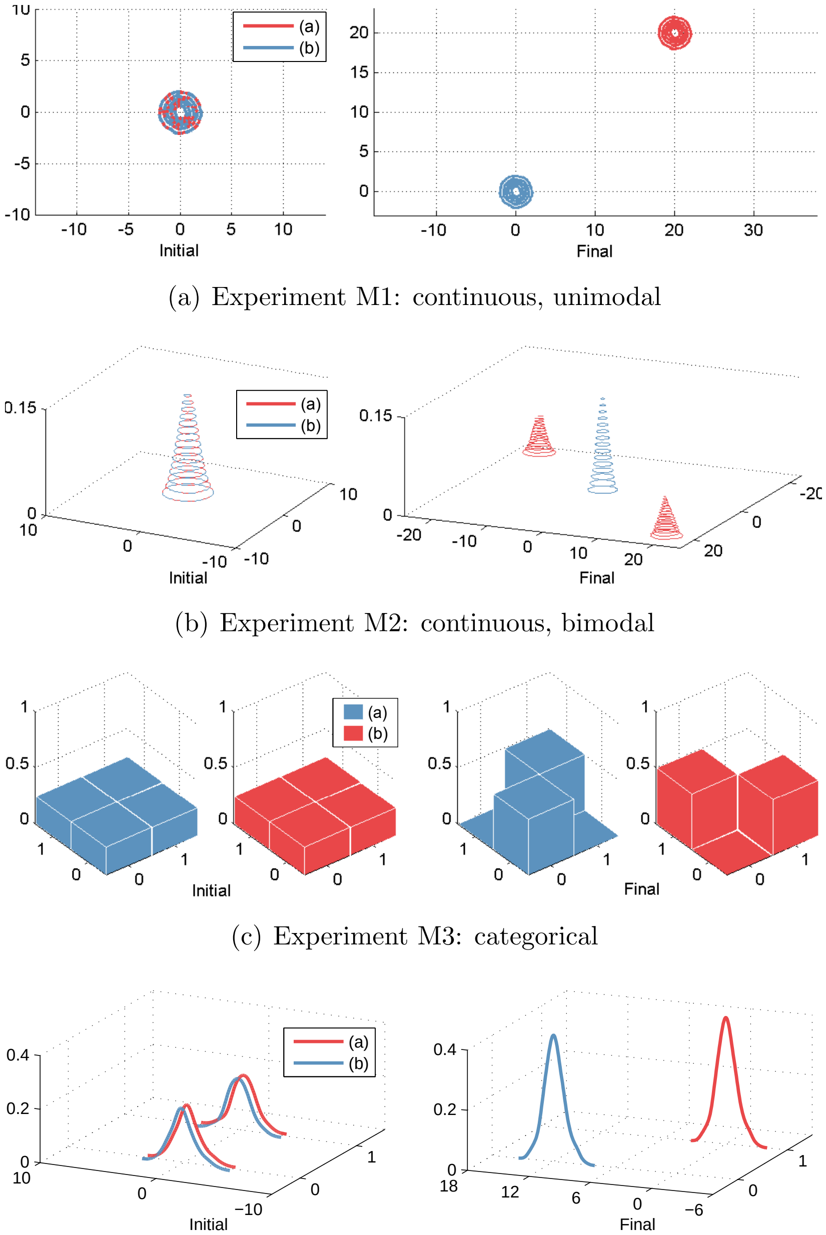
(b) Experiment U2: continuous, bimodal

(c) Experiment U3: continuous, non-parametric

(d) Experiment U4: binomial, ordinal

(e) Experiment U5: categorical, non-ordinal

Figure 3.1: Initial and final states of the probability distributions for the univariate experiments



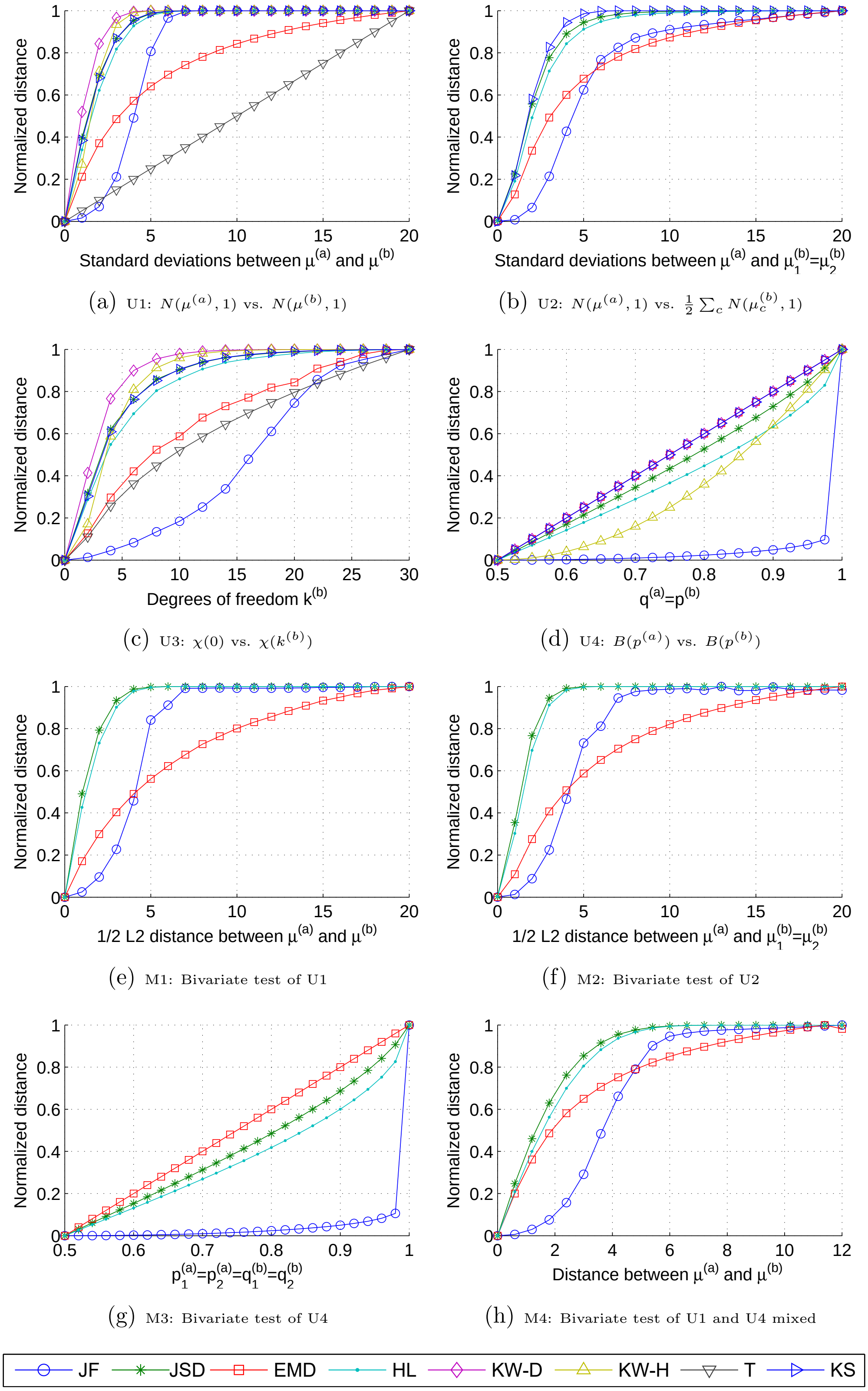
(a) Experiment M1: continuous, unimodal

(b) Experiment M2: continuous, bimodal

(c) Experiment M3: categorical

(d) Experiment M4: mixed continuous and categorical

Figure 3.2: Initial and final states of the probability distributions for the multivariate experiments



(a) U1: $N(\mu^{(a)},1)$ vs. $N(\mu^{(b)},1)$

(b) U2: $N(\mu^{(a)},1)$ vs. $\frac{1}{2}\sum_c N(\mu_c^{(b)},1)$

(c) U3: $\chi(0)$ vs. $\chi(k^{(b)})$

(d) U4: $B(p^{(a)})$ vs. $B(p^{(b)})$

(e) M1: Bivariate test of U1

(f) M2: Bivariate test of U2

(g) M3: Bivariate test of U4

(h) M4: Bivariate test of U1 and U4 mixed

Figure 3.3: Results of univariate, (a), (b), (c) and (d), and multivariate, (e), (f), (g) and (h) experiments. JF: Jeffrey, JSD: Jensen-Shannon, EMD: Earth Mover's Distance, HL: Hellinger, KW-D: Kruskal-Wallis mean rank difference, KW-H: Kruskal-Wallis statistic, T: t-test statistic, KS: Kolmogorov-Smirnov statistic.

Chapter 4

Multi-source variability metrics for biomedical data based on simplicial projections from probability distribution distances

Biomedical data may be composed of individuals generated from distinct, meaningful sources. Due to possible contextual biases in the processes that generate data, there may exist an undesirable and unexpected variability among the probability distribution functions of the source subsamples, which, when uncontrolled, may lead to inaccurate or unreproducible research results. Classical statistical methods may have difficulties to undercover such variabilities when dealing with multi-modal, multi-type, multi-variate data. This chapter proposes two metrics for the analysis of variability among multiple data sources, robust to the aforementioned conditions, and defined in the context of data quality assessment. Specifically, a global probabilistic deviation (GPD) and a source probabilistic outlyingness (SPO) metrics are proposed. The first provides a bounded degree of the global multi-source variability, designed as an estimator equivalent to the notion of normalized standard deviation of PDFs. The second provides a bounded degree of the dissimilarity of each source to a latent central distribution. The metrics are based on the projection of a simplex geometrical structure which, based on the conclusions from Chapter 3, is constructed from the Jensen-Shannon distances among the sources PDFs. The metrics have been evaluated and demonstrated their correct behaviour on a simulated benchmark and with real multi-source biomedical data using the UCI Heart Disease dataset.

The contents of this chapter were published in the journal publication by Sáez et al (2014b)—thesis contribution P3. The developed methods are included in the software contributions S1 and S3.

4.1 Introduction

Biomedical data may be generated from different sources. Multi-centre data repositories are a well-known example. Other examples include data generated from different users, or groups of data at different levels of granularity through a sensible hierarchy, e.g., a geographical location. Hereafter multi-source data is defined as data comprising individuals generated from distinct, meaningful, originating sources, belonging each individual to a single, clearly identified, source.

Compiling data from multiple sources may ensure a good sample representation from a broader and more representative population. In fact, obtaining a representative and significant sample is usually the objective of multi-centre studies (McMurry et al, 2013).

However, due to possible contextual biases in the processes that generate data, multi-source data may also entail an unexpected or undesired variability among its sources, which can lead to contradictory or unreproducible results (McMurry et al, 2013). As a consequence, two situations may arise: (1) data consumers do not consider such variability, leading their results to poor hypotheses, models, or wrong decisions; (2) data consumers are aware about the possible variability but the complexity of data either hinders such discovery or they do not have the proper discovery methods. Regarding to (2), Sáez et al (2013b) showed that classical statistical tests may have difficulties or be not suitable at all when dealing with specific data features, such as in multivariate, multi-type and multi-modal data. In any of the cases, one could perfectly draw hypothesis or obtain acceptable models from data assuming that data is stable among sources—i.e., modelling and evaluation made with data from all sources. However, it may not be assured that these results will either maintain the same effectiveness when used or evaluated at a single source or be generalisable at all to other sources.

This variability among sources is in fact a variability among their data probability distribution functions (PDFs)[c]. Ideally, biomedical research studies, such as clinical trials or population studies, would expect PDFs to be stable among the different sources in order to draw generalizable conclusions. However, if PDFs show variability among the data sources, data fail to meet user expectations what, by definition (Wang and Strong, 1996), results in a lack of data quality. The variability among sources has been addressed by some authors in the biomedical data quality domain (Cruz-Correia et al, 2010; Weiskopf and Weng, 2013). Nevertheless, it has mainly been related to semantic, structural or element agreement among sources. In this work, the variability among sources' PDFs is studied as a *multi-source variability* data quality dimension. Studying the variability among data sources may help data consumers understand their data, detect problematic or biased sources, detect patterns among the sources or, more generally, take better decisions in the research process.

In this work, a method for obtaining representative measurements of the data source variability is presented. It contributes to the state-of-the art with two metrics of a multi-source variability data quality dimension, designed as a descriptive statistical method to assess multi-source variability, and being robust to the aforementioned features where classical statistical tests may not be suitable. The first metric measures the degree of global multi-source variability—i.e. global probabilistic deviation (GPD)—and the second the degree of outlyingness of single sources—i.e., source probabilistic

[c] Note that semantic or structural consistence among data sources is not discussed here, which is out of the scope of this thesis.

outlyingness (SPO), being both designed to be comparable among different domains or datasets. The calculus of metrics is based on the projection of a D-dimensional simplex constructed from the pairwise PDF distances among sources. Additionally, this method provides the basis for a source clustering and for the spatial visualization of data source variability. The method is evaluated with simulated and real data using the UCI Heart Disease dataset (Asuncion and Newman, 2007; Detrano et al, 1989).

The rest of the paper is organized as follows. Section 5.2 reviews different variability problems found in biomedical studies, the statistical methods usually employed to detect such variabilities, and settles the work in the context of data quality. Section 4.3 describes the simplex geometrical structure and some of their properties. Section 5.3 describes the multi-source variability methods presented in this work. The experiments to evaluate the method and the results are described in Section 4.5. Finally, Sections 5.6 and 4.7 describe the discussion and conclusions of the work.

4.2 Background

4.2.1 Variability in biomedical data

The outcomes of biomedical research and healthcare practice depend on taking decisions based on the available information (Cruz-Correia et al, 2010). The data behind such information is registered by humans or devices based on observations of facts, at any stage of the healthcare process, and under an environment or context. As a consequence, the interpretation of such observations may be different according to different contexts. In addition, latent contexts (e.g., the socio-economic profile of a geographical location) can have a direct influence on the original facts, independent on its interpretation. In other words, contextual biases in the processes that generate data may have associated an undesired or unexpected variability among the data-generating sources.

Many examples in the literature can be used to illustrate these types of variabilities. Markus et al (1997) found differences in the interpretation of a common dataset of Doppler embolic signals among different centres, even using the same equipments. Verwey et al (2009) and Mattsson et al (2010) found diagnostic variabilities among centres in several multi-centre studies evaluating the use of cerebrospinal fluid biomarkers for Alzheimer's disease. Verwey et al. recommended the standardization of procedures and homogenization of assays to reduce such variability. Such reduction was proved by Dargaud et al (2012) in the use of a thrombin generation test in clinical trials. However, Pagani et al (2010) encountered that even using a common acquisition protocol, differences were still found among centres in diffusion tensor magnetic resonance imaging findings. On the other hand, as a single but relevant example of how the context can cause such variations, Jarman et al (1999) showed that some hospital characteristics have a direct interaction with the ratio of hospital death rates.

According to the type or purpose of the study, detecting and measuring multi-source variabilities are generally addressed by means of classical statistical methods. In clinical trials, the coefficient of variation or, its non-parametric equivalent, the quartile coefficient of dispersion are generally used to measure variabilities among some numerical indicators obtained from each source. These methods have some possible drawbacks. Summarizing in one scalar indicator the original distribution of what is measured on each source may in some cases entail an information loss. Whilst the coefficient of variation may be affected by the scale or type of the analysed variable (e.g., a mean near 0 on a non-ratio scale), the quartile coefficient of dispersion may miss additional information about the shape of variable PDF. One advantage of the quartile coefficient of dispersion is that is unit-free, and so is comparable among different problems.

Classical statistical tests used to contrast differences among two or more univariate data samples include One-way Analysis of Variance (ANOVA) for Gaussian data, Kruskal-Wallis test for non-Gaussian data, and χ^2 test for categorical. These tests are not designed to deal with multivariate or multi-type data. In addition, both the two-sample equivalent of One-way ANOVA, the Student's t-test, and the Kruskal-Wallis test have problems with multi-modal data, as shown in the previous section (Sáez et al, 2013b). Though, it is also expected in ANOVA, which is suited to unimodal and homoscedastic Gaussian data.

Another method to test differences on samples composed by numerical and categorical data is the N-way ANOVA. It evaluates the effect of multiple factors, the categorical variables, on a dependent numerical variable. Hence, it is not suited to measure the variability in the joint distributions of numerical and categorical variables.

Finally, the Multivariate ANOVA (MANOVA) test is suited when having more than one dependent variable. Analogous to One-way ANOVA, variables must be numerical, Gaussian and homoscedastic. While MANOVA may be useful under these assumptions, the contrast is made on linear combinations of the variables, where such a collinearity may not exist among these.

The multi-source variability metrics developed in this work are based on information-theoretic methods to measure PDF distances. As an alternative to classical statistical tests, information-theoretic methods are able provide more information about the variability between data distributions where the assumptions of the classical tests are not met (see Section 4.2.3).

The method presented in this work does not intend to replace the aforementioned tests for their specific use scenarios. Its purpose is to provide a metric for the variability among different sources of data and the degree of outlyingness of single sources, being (1) suitable to multivariate, multi-type and multi-modal data, and (2) bounded and therefore comparable among different problems. Additionally, it intends to (3) pose an alternative to the classical statistical tests for those cases where the aforementioned conditions of data hinder or impede their use.

4.2.2 Data source variability in the context of Data Quality

The variability among sources has been addressed by several authors as a data quality problem from different perspectives. Cruz-Correia et al (2010) reviewed different issues associated to data integration and sharing among different health information systems or organizations. They found structural and semantic interoperability as the major problems. Weiskopf and Weng (2013) carried out a systematic review on the methods and dimensions of data quality assessment in the context of reuse of electronic health records (EHRs) for research. From a set of 95 articles they derived five high-level dimensions and seven assessment methods. From these, the *concordance* dimension, and the *data source agreement* and *distribution comparison* methods can be related to our problem. They defined concordance as *Is there agreement between elements in the EHR, or between the EHR and another data source?*. Hence, concordance can refer to the agreement among observations of a patient EHR, agreement among the same observation of a patient on different information systems, or agreement among a set of EHRs with respect to a gold standard with the same information. Whilst the last two are related to the variability among sources, only the last is related to the problem of comparing data probability distributions. Though, they identified the method of comparison with a gold-standard distribution as a method to assess the concordance dimension. However, any of the articles comprising the systematic review neither intend to provide a variability metric among a set of sources nor put attention on the heterogeneous features of biomedical data.

4.2.3 Dissimilarities between biomedical data distributions

Biomedical data usually show heterogeneous conditions. Concretely, biomedical data are generally 1) multivariate (i.e., data have more than one variable), 2) multi-type (i.e., simultaneously continuous, discrete ordinal and non-ordinal variables), and 3) multi-modal (i.e., data distributions are generated by more than one mode). In Chapter 3, we studied the behaviour of different PDF dissimilarity metrics envisaging these data features. The results of such study are summarized in Table 3.1.

The results showed that the aforementioned data features may complicate the application of classical statistical or data analysis methods for the assessment of differences among data samples. Specifically, the results confirmed that classical statistical tests may have difficulties on multi-modal data, or may not be not suitable at all on multivariate or multi-type data. Information-theoretic distances, including the Jeffrey and Jensen-Shannon distances, and the Earth Mover's Distance resulted the most suitable distances to all conditions. Information-theoretic are distances which derive from the Shannon's entropy theory, while EMD derives from the digital imaging field as a measure to calculate the minimum cost of transforming one histogram into another.

Regarding to the information-theoretic distances, when the probability mass in any region of the support in any of the compared distributions tends to zero, the Jeffrey distance (symmetrized version of Kullback-Leibler divergence) tends to infinite. In contrast, the Jensen-Shannon distance (JSD) is a metric bounded between zero and one, and it was smoothly convergent to one on that situation. In fact, such bounds facilitate the distance comparison on different problems. As a consequence, and although both EMD and JSD could be suitable for the purpose of the multi-source variability method, the JSD was chosen for its direct foundations in information-theory and for its generalization for comparability.

4.3 Simplices and properties

Generally speaking, a simplex is the generalization of a triangle to D dimensions, $D \in \mathbb{N}$. A D-simplex, Δ^D, is composed by $v_1, ..., v_n : n = D + 1$ vertices, which form the convex hull of the simplest polytope in R^D. Simplices can be regular or irregular. Some properties of these that will be required in the development of the variability metrics are described next.

A simplex is regular when the distances among their vertices are equal. Consequently, the length of the segment formed from the centroid of the simplex to each vertex is also equal. The angle γ between any pair of these segments depends on the number of dimensions and is (Parks and Wills, 2002):

$$\gamma(D) = \arccos(\,^{-1}/_{D}) \tag{4.1}$$

The simplex when all the distances between its vertices are one will be defined further on as 1-regular (1R) simplex. In any D, any pair of vertices and the centroid of the simplex form a triangle. Thus, according to the *law of sines*, the distance $d(v, O) = d_{1R}(D)$ between any vertex and the centroid on 1-regular simplices in D dimensions is defined as:

$$d_{1R}(D) = \frac{1}{2\sin(^{\gamma(D)}/_{2})}, \tag{4.2}$$

where $d_{1R}(1) = {}^1/_2$ as a continuity convention in $D = 1$ (two vertices). See Section B.1 for details.

On the other hand, a simplex is irregular when at least one of its vertices is at a different distance from the centroid with respect to the others. Consequently, the distances between vertices do not have to be equal. In that case, if it is defined as a simplicial space upper-bounded by a 1-regular simplex—i.e., the simplicial space containing all the possible simplices where the maximum distance among vertices is one—, the distance of any vertex to the centroid of the irregular simplex will be bounded by:

$$d_{max}(D) = 1 - \frac{1}{D+1}, \tag{4.3}$$

which is larger than $d_{1R}(D)$ for the same D. See Section B.2 for details.

4.4 Methods

The multi-source variability method provides two metrics of the data source variability: 1) the global probabilistic deviation (GPD—Ω), and 2) the source probabilistic outlyingness (SPO—$\mathbb{O}$). The GPD measures the degree of global multi-source variability. The SPO of a single source is understood as a measure of the distance of its PDF to a latent central distribution of all the sources. These metrics are obtained based on the simplex where each vertex represents a data source, and its edge lengths the pairwise PDF distance between the data of the sources represented by the adjacent vertices. A multi-source variability plot for the visualization of the data source variability can be derived as a by-product of the process. Figure 4.1 shows the procedure to obtain these outcomes. In the rest of the section, the different steps of the procedure are described. The procedure input is a multi-source dataset $X = (X_1, ..., X_S)$, where X_s is the sub-sample of data corresponding to source s and S is the total number of sources.

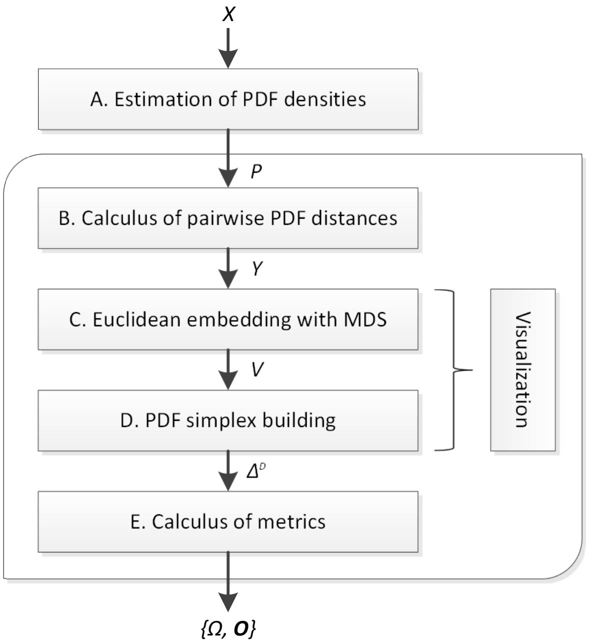

X
A. Estimation of PDF densities
P
B. Calculus of pairwise PDF distances
Y
C. Euclidean embedding with MDS
V
D. PDF simplex building
Δ^D
E. Calculus of metrics
Visualization
{Ω, O}

Figure 4.1: Steps of the method to obtain the multi-source variability metrics: global probabilistic deviation (GPD—Ω), and source probabilistic outlyingness (SPO—$\mathbb{O}$). Each step is described in its corresponding subsection in Section 5.3.

4.4.1 Estimation of PDF densities

The objective of this step is to obtain the set P of representative PDFs of the data of each data source as $P = (P_1, ..., P_S), P_s : p(X_s)$. Depending on the characteristics of the data or the problem, different preprocessing or density estimation methods may be chosen. In low dimensional problems, histograms or, a smoothing method for the numerical case, Kernel Density Estimations (see Section 2.2.1) may be used. In higher dimensional problems data can be embedded into a lower-dimensional representation using dimensionality reduction methods such as Principal Component Analysis (PCA) or non-linear manifold embeddings, such as ISOMAP (Tenenbaum et al, 2000). As a consequence, the estimation of the probability distribution functions becomes easier, being as much of the original information conserved. Depending on the layout of data it is important to choose between linear and non-linear dimensionality reduction methods, as linear methods (such as PCA) may fail on projecting non-linear continuities of data points on the original higher-dimensional space. In addition, in the mixed multi-type case, i.e. when numerical and non-ordinal categorical variables coexist, a special density estimation may be required when histograms become noisy or sparse. In that case, a solution may be obtained using non-linear dimensionality reduction methods which allow defining a distance metric among the values of categorical data (e.g., using

ISOMAP). In any case, this stage of the method is flexible to the use of different density estimation methods, thus, the selection of the proper density estimation method is out of the scope of this work.

The output of this step is the set P of PDFs with:

$$P = (P_1, ..., P_S), P_s : p(X_s) \tag{4.4}$$

4.4.2 Calculus of pairwise PDF distances

In this step the pairwise PDF distances among all sources are calculated. These distances correspond to the magnitude of the edges of the simplex under construction. Hence, being S the number of sources, the number of distances to be calculated and, therefore, the number of edges of the simplex, corresponds to the binomial coefficient $\binom{S}{2} = \frac{S!}{2!(S-2)!}$.

As discussed in Section 4.2.3, and according to the results of chapter 3, the pairwise PDF distance $d(P_s, P_{s'})$ between PDFs P_s and $P_{s'}$ will be calculated based on the Jensen-Shannon distance:

$$d(P_s, P_{s'}) = JSD(P_s||P_{s'}), \tag{4.5}$$

where $JSD(P_s||P_{s'})$ is the Jensen-Shannon distance in Equation 2.36, which defined in the $[0, 1]$ interval when using the base 2 logarithm to calculate the Kullback-Leibler divergence (Equation 2.29).

The discrete Kullback-Leibler divergence in Equation 2.29 allows computing the non-parametric Jensen-Shannon divergence on D-dimensional histograms by computing for each bin in the common support the corresponding discrete Kullback-Leibler summations. However, the Jensen-Shannon divergence can also be calculated analytically based on the parametric, analytical forms of the Kullback-Leibler divergence (see Section 2.2.2).

The output of this step is the S-by-S symmetric dissimilarity matrix Y:

$$Y = (Y_{11}, ..., Y_{SS}), Y_{ss'} : d(P_s, P_{s'}) \tag{4.6}$$

4.4.3 Euclidean embedding using multidimensional scaling

The Information Geometry field states that probability distributions lie on a Riemannian manifold which inner product is given by the Fisher information metric corresponding to a specific family of distributions (Amari and Nagaoka, 2007). The geodesic distance between the points representing probability distributions in such a statistical manifold can be approximated by means of PDF distances, such as the Jensen-Shannon. In this work, only the distances among a set of distributions are known. They are not restricted to a specific family, hence, it can be considered that they lie on a statistical manifold of unknown configuration (i.e., inner product and thus dimensionality). To the purpose of this work, a simplex must be constructed from such probabilistic distances in a $\mathbb{R}^D$ space. To this end, MDS is used, which calculates an Euclidean embedding of a inter-point dissimilarity matrix—see Section 2.2.4.

Hence, given the dissimilarity matrix $Y = (Y_{11}, ..., Y_{SS})$, MDS will obtain the $V = (V_{11}, ..., V_{SD})$ coordinates for the S sources' PDFs in a $\mathbb{R}^D$ Euclidean space.

To the purpose of the multi-source variability metrics, the PDF dissimilarities should be approximated as better as possible, while maintaining the $[0, 1]$-bounds. Full-dimensional scaling provides a perfect embedding when the input dissimilarities are Euclidean, however, this is not ensured for all types of dissimilarities, such as the Jensen-Shannon distance. However, as mentioned in Section 2.2.4, Riemannian manifolds can be isometrically embedded into an Euclidean space for a sufficient number of dimensions. In this case, were as using full-dimensional scaling, which provides the maximum dimensional embedding, facilitating the isometric change of space. In fact, in all the embeddings carried out for the evaluation of this work (Section 4.5), a zero Stress was obtained. As a consequence, we can conclude that maintaining the $[0, 1]$-bounds cannot be considered an issue in this work.

The output of this step is the S-by-D coordinates matrix V:

$$V = (V_{11}, ..., V_{SD}), \tag{4.7}$$

where V_{sd} is the d^{th} significant coordinate of source s.

4.4.4 PDF simplex building

Each of the points obtained in the previous step represents a source PDF, and the euclidean distances among them keep the corresponding pairwise PDF distance. These S points and the $\binom{S}{2}$ edges represent the vertices and edges of a D-dimensional simplex. This simplex and its centroid stand as the basis of the proposed method.

Given that the pairwise PDF distances are upper limited by one, the distances between vertices are so. It makes the corresponding simplicial projection meeting the next properties:

Property 1 *For a specific number of sources $S = D + 1$, whatever the PDF distances among them, the maximum possible simplicial projection (i.e., when the distances of all vertices to the centroid are maximum) is a D-dimensional 1R simplex.*

Property 2 *In the case of a D-dimensional 1R simplex, the maximum distance between any vertex and the simplex centroid is $d_{1R}(D)$ (Equation 4.2).*

Property 3 *In the general D-dimensional case (irregular simplices) the maximum distance between any vertex and the simplex centroid will be bounded by $d_{max}(D)$ (Equation 4.3).*

Properties 1 and 2 establish the theoretical maximum inter-source dissimilarity state, thus defining an upper bound of global multi-source variability. It is straightforward that the lower bound occurs when all distributions are equal and thus all points are the same. On the other hand, in Property 3, $d_{max}(D)$ establishes the limit for the cases where $d(P_s, P_s') = 0 : s, s' \in \{1, ..., S-1\}$ and $d(P_s, P_S) = 1$ (the distance among all the sources except one is 0, and the distances between this one and the formers are 1).

The output of this step is the D-dimensional simplex Δ^D:

$$\Delta^D = (V, C), \tag{4.8}$$

where V correspond to the coordinates of the vertices and and C to the simplex centroid (Equation 4.9), both defined in $\mathbb{R}^D$.

$$C = \sum_{s=1}^{N} \frac{V_s}{N}, \tag{4.9}$$

4.4.5 Calculus of metrics

The purpose of this step is to calculate the GPD and the SPO metrics based on the simplex obtained in the previous step. This simplex represents a projection of the sources' PDFs keeping the dissimilarities among them. As a consequence, it can be affirmed that the simplex centroid may represent a latent central point with respect to all PDFs, and two definitions can be derived:

Definition 1 *The centroid C of Δ^D represents a latent central tendency of the original measured population.*

Definition 2 *The distance of a vertex V_s to the centroid C, $d(V_s, C)$, represents the deviation of a data source with respect to the central tendency of the population.*

As a consequence, the closer the PDFs vertices are to the centroid, the more stable the dataset is, while the larger the more unstable. The resultant simplex is bounded by an 1R simplex, as described in the previous step. Additionally, the larger the distance of a vertex from the centroid, the more outlying a source is with respect to the latent central tendency. The variability metrics proposed on this work are based on such definitions.

Global probabilistic deviation

The standard deviation is a measure of the variability of a sample with respect to its central tendency. If the sources' PDFs are considered as individuals of a population and the centroid as its central tendency, the notion of standard deviation can be directly applied to obtain a measure of the variability of the PDFs. In fact, as the PDF points are embedded in a $\mathbb{R}^D$ Euclidean space where the triangle inequality holds, their distances to the centroid can be considered as PDF distances to a latent central distribution. Hence, the derived standard deviation among S PDFs can be defined as:

$$Std(P_1, ..., P_s) = \frac{\sum_{s=1}^{S} d(V_s, C)}{S}, \tag{4.10}$$

where $d(V_s, C)$ is the Euclidean distance between the vertex V_s and the centroid C. Note that as distances are always positive, the resultant deviation is given in the original units.

However, despite the pairwise PDF distances are $[0,1]$-bounded independently of the number of dimensions D, the distances between each vertex and the centroid are neither defined in the same space for different D nor $[0,1]$-bounded. It causes the standard deviation measurement of Equation 4.10 neither to be comparable when the number of sources S (and therefore D) is different, nor $[0,1]$-bounded. This situation would not permit the deviation to be comparable among different domains. Using Property 2 of Section 4.4.4, the solution comes by normalizing the standard deviation by the maximum deviation on D dimensions given the upper multi-source variability bound, i.e. $d_{1R}(D)$. In fact, that upper bound distance is the upper bound of the standard deviation on D, with $Std(P_1,...,P_s) \rightarrow d_{1R}(D)$ in irregular simplices, and $Std(P_1,...,P_s) = d_{1R}(D)$ in the case of 1R simplices. That makes the measurement comparable and upper bounded to one, and leads to the definition of the GPD metric:

Definition 3 *The global probabilistic deviation metric Ω among a set of datasets $X = (X_1,...,X_S)$ is defined as:*

$$\Omega(X_1,...,X_S) = \frac{Std(P_1,...,P_s)}{d_{1R}(D)} \tag{4.11}$$

Source probabilistic outlyingness

The distance $d(V_s, C)$ gives a degree of how far a source is from the central tendency of the population (Definition 2). However, as well as in the GPD metric, that distance is defined in different spaces according to D, thus making the distance neither comparable nor $[0,1]$-bounded. Analogously to the GPD metric, a normalization factor is required. In this case it is the distance between a single vertex and the centroid what must be normalized. Hence, using Property 3, the normalization factor is given by $d_{max}(D)$, leading to the definition of the outlyingness metric:

Definition 4 *The source probabilistic outlyingness metric $\mathbb{O}$ of a dataset X_s with respect to the central tendency among the datasets $X_1,...,X_S$ is defined as:*

$$\mathbb{O}(X_s) = \frac{d(V_s, C)}{d_{max}(D)} \tag{4.12}$$

4.4.6 Multi-source variability (MSV) plot

Although the objective of this work is to provide metrics for the data multi-source variability, it must be mentioned that this method also provides the means to visualize the variability or interdependences among data sources. In fact, the visualization of complex scientific datasets using aggregated data is of special research interest (Wong et al, 2000).

Concretely, the simplex coordinates calculated by MDS serve as a D-dimensional visualization of the multi-source variability, where the d^{th} coordinate is the d^{st} important in terms of conserving the real distance. Due to the obvious restriction that visualizations can be provided up to three dimensions, the most accurate visualization is obtained taking the first two or three simplex coordinates. In the next sections some examples of a basic MSV plot are provided. Next, in Chapter 6 an improved MSV plot is provided, where each data source is presented as a circle which radius is made proportional to the number of cases in that data source, and the color of each circle indicates the SPO of the source.

4.5 Evaluation

The variability metrics presented in this work have been first evaluated for scalability on different simulated conditions. Second, real multi-source biomedical data have been used with the purpose of completing the evaluation on real data variables and compare results with other classical statistical methods. In this section the evaluation experiments and their results are presented.

4.5.1 Evaluation of scalability

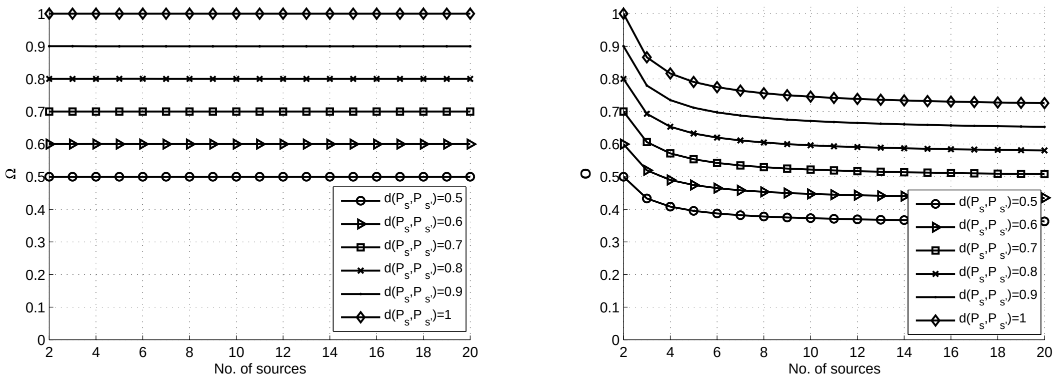
In this evaluation the GPD (Ω) and the SPO ($\mathbb{O}$) metrics were tested for scalability against variations in the number of sources, variables, and distributional dissimilarities. The GPD and SPO were measured and plotted at each iteration. Using the Jensen-Shannon distance in combination with non-parametric PDF estimations, the variability metrics are constructed to be robust against different variable types and multi-modality as shown in Chapter 3. As a consequence, to simplify the interpretation of these experiments unimodal Gaussian variables and analytical parametric Jensen-Shannon distances were used. We recall that the analytical measurement of distances between the Gaussian variables is made based on their parameters, therefore, the experiments did not require generating random data individuals.

Different number of sources

New data sources were iteratively added at the same pairwise distance with respect to the previous sources. This leads to regular simplicial projections, thus, the SPO is the same for all sources at each iteration. Measurements were taken for different source pairwise distances. Results are shown in Figure 4.2.

The GPD metric keeps stable as the number of sources increases. This stability is a consequence of the normalization of the metric by $d_{1R}(D)$. This normalization leads to an additional interesting property of the GPD metric, by which in the case all pairwise distances are the same the metric is equivalent to that distance. Additionally, in the case all sources are at the maximum pairwise distance, i.e., one, the GPD is bounded to one as well.

On the other hand, the outlyingness metric shows a non-linear negative tendency which converges in all pairwise distances. As the number of sources at the same pairwise distance increases, the distance of vertices to the centroid does so until convergence. However, according to Property 3, in the case that pairwise distances are not the same among all sources, i.e. an irregular simplex, an independent source may be at a larger distance from the centroid than in the regular maximum case. Such irregular maximum corresponds to the normalization factor for outlyigness. Hence, as



(a) Global probabilistic deviation (Ω)

(b) Source probabilistic outlyingness ($\mathbb{O}$)

Figure 4.2: Results on different number of sources. Measurements were taken for different inter-source pairwise PDF distances, given by $d(P_s, P_{s'})$. The GPD keeps stable as the number of sources increases (a), while the SPO converges as the number of sources increases (b).

an expected property of the metric, when a source is at a large distance to a group of sources which are close among each other, the former will be more likely an outlier when the number of sources in the latter group increases.

Different number of variables

Given two multivariate Gaussian sources their number of variables is increased. The means of the first variable are at a fixed distance between the two sources, while the rest of the variables are equal (covariance matrices were diagonal with $\Sigma_{ij} = 1$). Hence, the purpose is to evaluate whether the variability caused by the first variable is maintained as new variables are included. Measurements were taken for different mean distances. Results are shown in Figure 4.3, where, as in the case of two sources the GPD and SPO are equivalent, only one plot series is shown representing both.

Results show the scalability of the metrics with the number of variables, as metrics keep stable as the number of variables increases. Hence, given a dissimilarity in a variable subspace, both GPD and SPO will theoretically be stable independently of the size of the full variable space.

Irregular source dissimilarities

In the general case differences among data sources will be irregular. That is, some sources may be close to each other, while others may show a higher outlyingness due, e.g., to sample biases. In this test this situation was evaluated. Using three bivariate Gaussian data sources with equal and diagonal covariance matrices, their means were iteratively and irregularly separated starting from an equal state until a convergence of the variability metrics. Concretely, sources 1 and 2 were smoothly separated from each other while source 3 equally separated from both with a larger velocity, expecting a larger outlyingness on it. Results are shown in Figure 4.4.

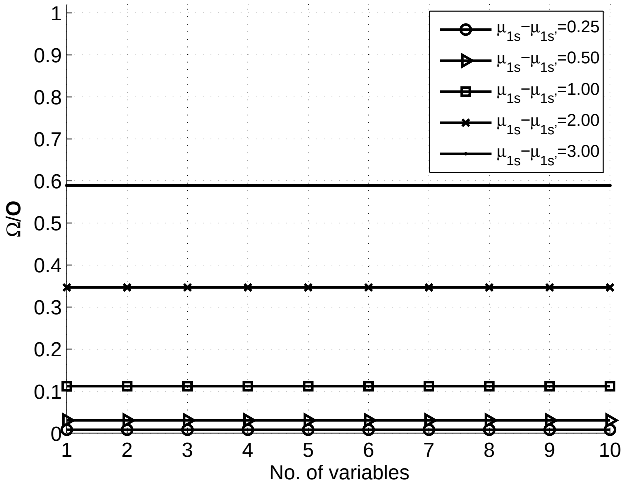
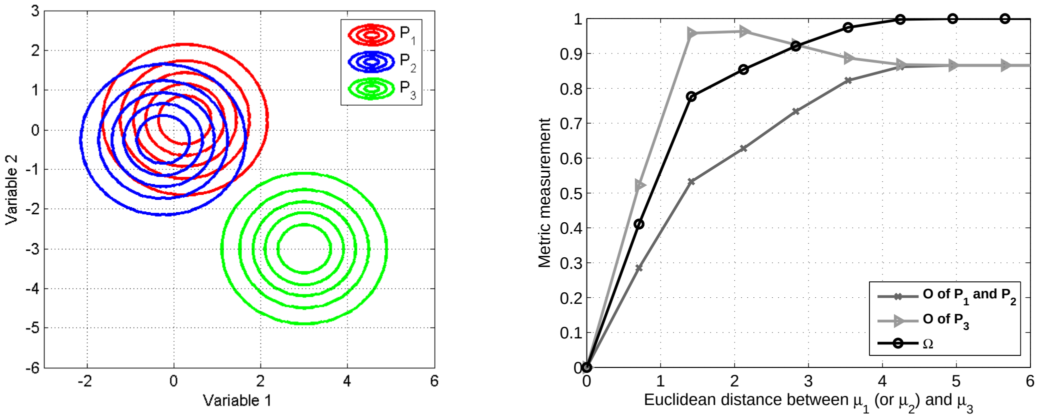


Figure 4.3: Results on different number of variables, where $\mu_{1s} - \mu_{1s'}$ indicates the Euclidean distance between means of the first variable of sources s and s'.



(a) Compared distributions in a intermediate iteration.

(b) GPD (Ω) and SPO ($\mathbb{O}$) of the distributions. The SPO is equivalent for distributions P_1 and P_2.

Figure 4.4: Results on a iterative irregular inter-source separation.

Figure 4.4(b) shows the variability metrics obtained during the iterative source separation, where figure 4.4(a) illustrates the PDFs in an intermediate state of the evaluation. It can be observed that as sources separate each other, the GPD does so until convergence, as well as the SPO metric of each source. Regarding to the source outlyingness, P_1 and P_2 are always at the same distance to the simplex centroid, hence showing the same outlyingness. However, as P_3 is separated at a larger velocity it gets to large distance to the centroid which, once P_1 and P_2 have also achieved a larger probabilistic pairwise distance, is reduced. This is due to the repositioning of the simplex centroid, related to the increase of the edge length between P_1 and P_2, associated to their bounded PDF distance.

4.5.2 Evaluation on real data (UCI Heart Disease)

The UCI Heart Disease (Asuncion and Newman, 2007; Detrano et al, 1989) is a publicly available multi-source dataset concerning heart disease diagnosis. It contains 76 variables acquired at four different healthcare locations namely the Cleveland Clinic Foundation, OH; the Hungarian Institute of Cardiology, Budapest; the University Hospital, Zurich, Switzerland; and the V.A. Medical Center, Long Beach, CA.

Only 14 of the variables are actually used in research studies, seven numerical and seven categorical. To facilitate the evaluation of this work, data has been cleansed to remove missing data while keeping the maximum possible number of non-missing variables and individuals. This process is described in Table 4.1. Although in general only the Cleveland sub-dataset is used in research experiments due to its higher quality and number of individuals, in these experiments all datasets have been used with the purpose to assess the variability among all the sources.

	Original (14 variables)		Cleansed (11 variables)	
Source	Individuals	Total missing values	Individuals	Total missing values
Cleveland	303	6	303	0
Hungarian	294	782	261	0
Switzerland	123	284	45	0
VA	200	699	129	0
All	920	1771	738	0

Table 4.1: Data cleansing of the UCI Heart Disease dataset carried out in this work

The variability metrics have been evaluated on this dataset as follows. First they have been univariately measured, in both numerical and categorical variables, comparing the results with classical statistical univariate tests. Second, they have been measured for each combination of variables, containing pairs of numerical, categorical and mixing types. Finally, the variability metrics have been measured using all the variables.

For this evaluation, the discrete Jensen-Shannon distance (Equations 2.36 and 2.29) was used as the reference PDF distance. In the case of numerical variables, their corresponding discrete PDFs were obtained from their KDE estimations using MATLAB© (Ihler, A. and Mandel, M., 2003). Gaussian kernels and automatic bandwidth selection (Silverman, 1986) were used.

Univariate evaluation

For each variable, the GPD and SPO metrics were measured. Additionally, depending on whether the variable was numerical or categorical the classical ANOVA and χ^2

tests were performed reporting the corresponding p-values. Note that in the numerical case the ANOVA makes the assumption that variables are unimodal Gaussians, what may not be true. Results are shown in Table 4.2, which have been ordered by their GPD. Additionally, Figures 4.6, 4.7 and 4.8 show the probability distributions and 2D simplicial projections of the different variables.

		p-value		SPO ($\mathbb{O}$)			
Variable	GPD (Ω)	ANOVA	χ^2	Cleveland	Hungarian	Switzerland	V.A.
trestbps	.1156	.3001	-	.0908	.0733	.1174	.0959
fbs	.1550	-	3e-10	.0219	.1364	.1048	.2431
exang	.2228	-	8e-13	.1768	.1871	.1609	.2031
sex	.2299	-	2e-10	.2201	.1549	.1562	.2195
cp	.2827	-	1e-16	.1895	.3016	.2563	.1759
age	.3054	6e-37	-	.0863	.4426	.1433	.3252
thalach	.3642	5e-37	-	.3897	.2019	.3497	.2480
restecg	.3709	-	2e-56	.4847	.2725	.1668	.2874
oldpeak	.4635	4e-10	-	.3377	.3912	.3924	.3925
num	.6302	2e-38	-	.4491	.6203	.5528	.4360
chol	.6737	2e-92	-	.4030	.3915	.9706	.4353

Table 4.2: Results of univariate evaluation on the UCI Heart Disease dataset. The variability and outlyigness measurements (columns) are shown for each variable (rows). Variables are sorted by the their GPD metric. The ANOVA or χ^2 p-value is shown according to whether the variable is numerical or categorical.

It can be observed that the GPD metric and the p-values of statistical tests are in general inversely proportional (Spearman correlation of $-.7182$, combining ANOVA and χ^2 p-values), i.e. the larger the GPD measurement the more significant the differences are found by the tests. This reinforces the consistence of the metric, which in addition shows its independence with respect to the type of variable. However, such correlation must be interpreted with caution. First, the behaviour of p-values do not need to be linear, and depends on the number of individuals or outliers (see Figure 4.5 for further details). As an example, the *trestbps* variable, shows a large p-value. As it can be observed (Figure 4.6(a)), its PDFs are quite similar except an outlier in the V.A. sample. Removing such outlying individual largely reduces the p-value to .1272, while the GPD and the V.A. SPO are only reduced to .1062 and .0739, respectively. On the other hand, statistical tests may not be accurate on multi-modal distributions, where the variability metrics are robust. Such problem can be observed in the *oldpeak* variable 4.8(a)), where ANOVA provides a p-value larger than its numerical predecessors.

The results also show how outlying sources can be identified by the SPO metric.

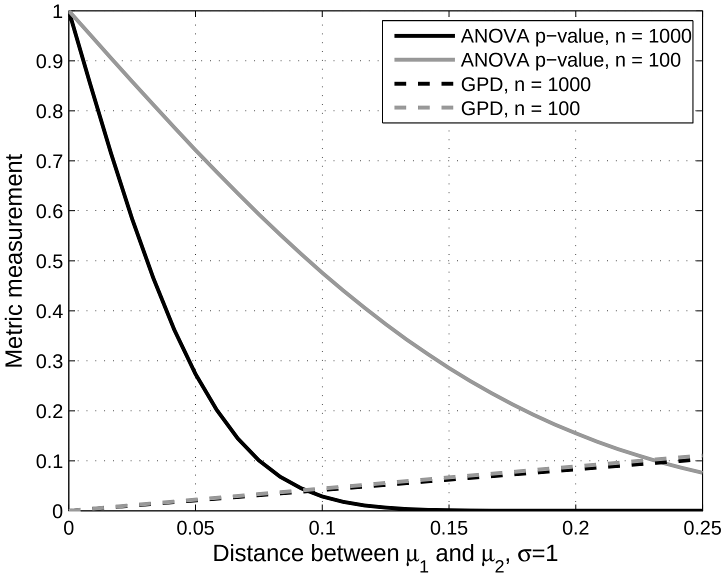
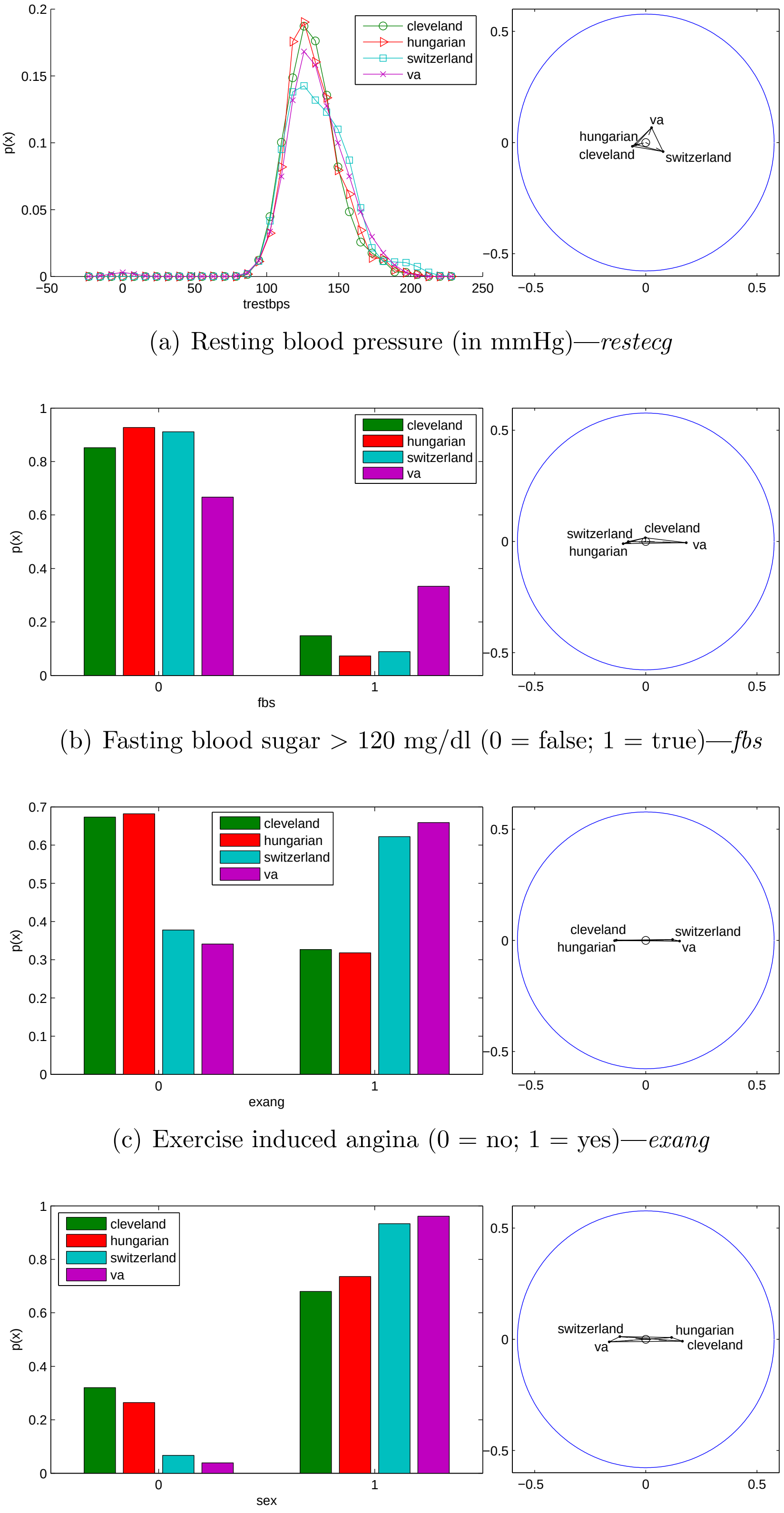


Figure 4.5: Comparison of the behaviour of the ANOVA p-value and the GPD (Ω) with different number of individuals. Two simulated Gaussian distributions with equal standard deviation were incrementally separated, where n random points were generated in each case. Probability density functions for GPD were estimated using KDE.

First, in the *age* variable, the respectively younger and older patients of Hungarian and V.A. datasets have their effect on their SPO metrics (Figure 4.7(b)). Regarding to the *chol* (serum cholesterol) variable, the Switzerland dataset showed an extreme outlyingness, probably caused by a wrong codification of the missing values: while in the Heart Disease dataset missing values are coded with -9, these seem to be coded with 0 (Figure 4.8(c)). In the *thalach* (maximum heart rate achieved) variable the projection shows the dissimilarity found among all sources (Figure 4.7(c)). Finally, the *num* variable corresponds to the heart disease diagnosis, and is the dependent variable for the data mining purposes of the dataset (note that studies with the Heart Disease dataset generally group positive values into a single positive class). However, it can be observed that there are large differences among the datasets. Specifically, the Hungarian dataset do not have patients with a value larger than 1, and Switzerland has very few healthy patients (0 value) in comparison with the others (Figure 4.8(b)).



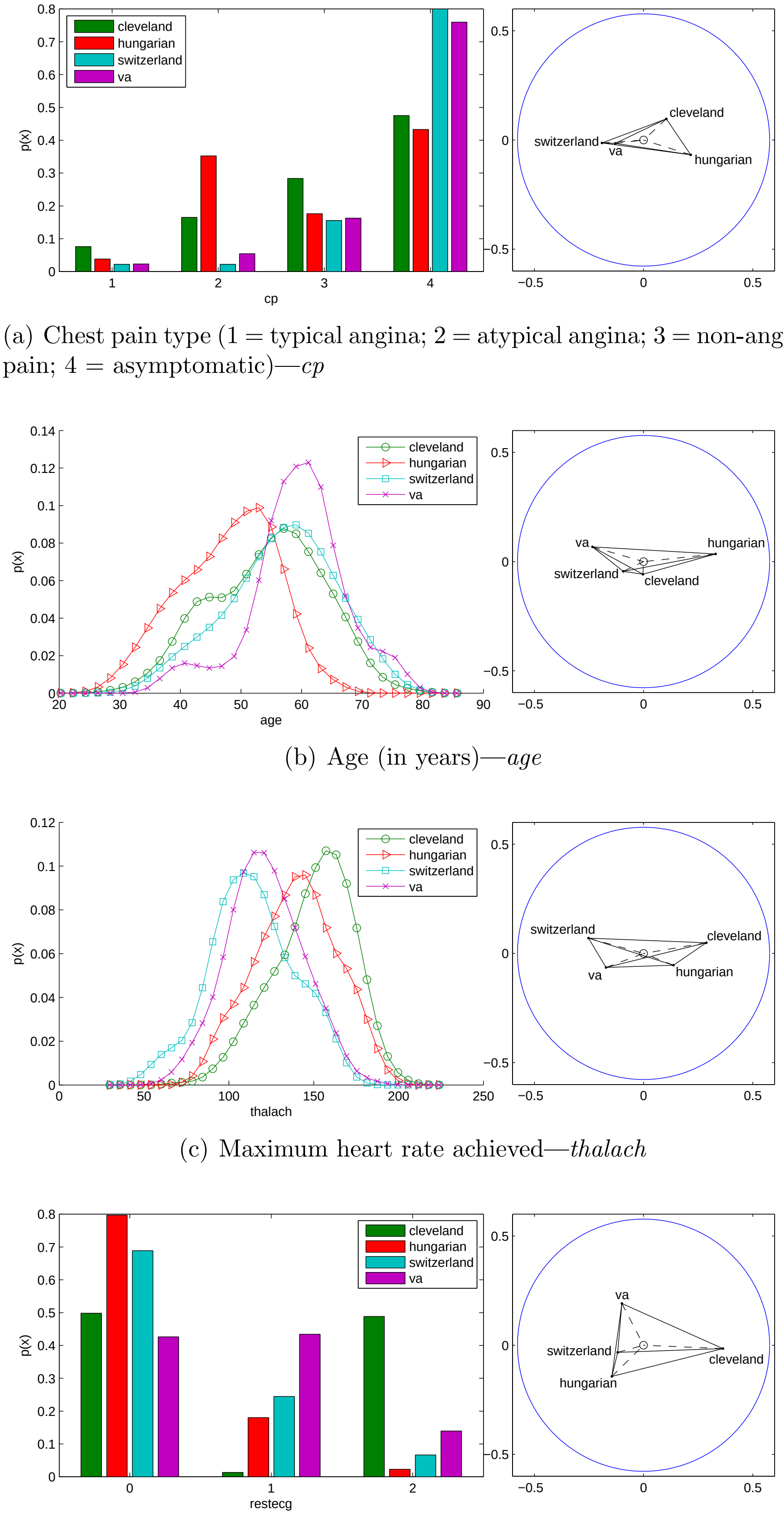
(a) Resting blood pressure (in mmHg)—*restecg*

(b) Fasting blood sugar > 120 mg/dl (0 = false; 1 = true)—*fbs*

(c) Exercise induced angina (0 = no; 1 = yes)—*exang*

(d) Sex (0 = female; 1 = male)—*sex*

Figure 4.6: Univariate probability distributions and 2-simplex variability plots for variables *trestbps*, *fbs*, *exang* and *sex*. The 2-dimensional sphere represents the upper variability bound where all the pairwise dissimilarities would be maximum.



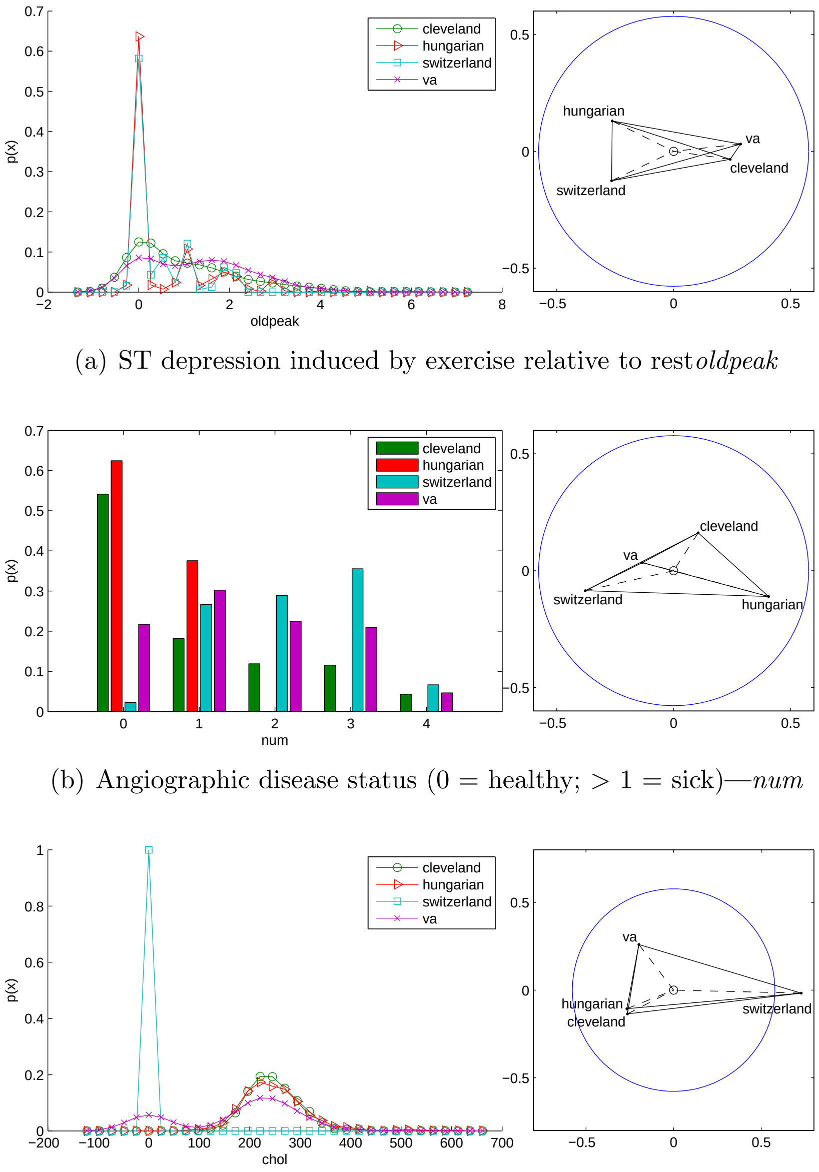
(a) Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)—*cp*

(b) Age (in years)—*age*

(c) Maximum heart rate achieved—*thalach*

(d) Resting electrocardiographic results (0 = normal; 1 = ST-T wave abnormality; 2 = left ventricular hypertrophy)—*restecg*

Figure 4.7: Univariate probability distributions and 2-simplex variability plots for variables *cp*, *age*, *thalach* and *restecg*. The 2-dimensional sphere represents the upper variability bound where all the pairwise dissimilarities would be maximum.



(a) ST depression induced by exercise relative to rest*oldpeak*

(b) Angiographic disease status (0 = healthy; > 1 = sick)—*num*

(c) Serum cholesterol (in mg/dl)—*chol*

Figure 4.8: Univariate probability distributions and 2-simplex variability plots for variables *oldpeak*, *num* and *chol*. The 2-dimensional sphere represents the upper variability bound where all the pairwise dissimilarities would be maximum.

Bivariate evaluation

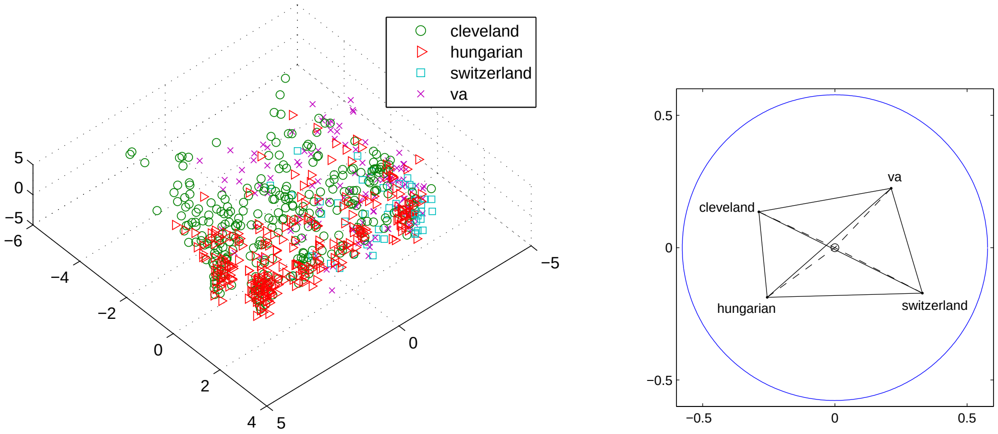
Results of bivariate evaluation are shown in Table 4.3. As described in 4.5.1, a low number of individuals makes histograms or density estimations to be more noisy due to data sparsity, thus, the low number of individuals on the evaluated dataset makes the GPD metric to tend being slightly higher in this bivariate test. However, these measurements are comparable among them, which permits discovering interactions of pair of variables (concretely of their joint probability) with respect to the data source. It can be observed that the large univariate variability of *chol* is reflected in all of its joint GPDs. On the other hand, the combinations including the dependent variable,*num* in this case, should take special attention by researchers as variability may indicate possible conflicts when developing predictive models based on the multiple datasets.

Variables	*sex*	*cp*	*trestbps*	*chol*	*fbs*	*restecg*	*thalach*	*exang*	*oldpeak*	*num*
age	.4123	.4515	.3516	.7562	.3416	.4992	.4917	.3999	.4006	.5469
sex	-	.3622	.2871	.7084	.2939	.4392	.4163	.2995	.5197	.6456
cp	-	-	.3687	.7160	.3550	.4939	.4703	.3344	.5568	.6714
trestbps	-	-	-	.6893	.2125	.4194	.3988	.2683	.2927	.4945
chol	-	-	-	-	.7005	.8357	.7367	.7065	.7080	.7797
fbs	-	-	-	-	-	.4138	.4198	.2947	.5042	.5928
restecg	-	-	-	-	-	-	.5287	.4420	.5950	.7063
thalach	-	-	-	-	-	-	-	.4022	.4580	.5789
exang	-	-	-	-	-	-	-	-	.4919	.6277
oldpeak	-	-	-	-	-	-	-	-	-	.5512

Table 4.3: Results of bivariate evaluation on the UCI Heart Disease dataset. Each cell shows the GPD (Ω) of the joint probability of the variables in the corresponding row and column.

Multivariate evaluation

The variability metrics were measured using all the available variables to assess the general variability of the complete dataset. To illustrate this example the PCA dimensionality reduction method with dummy coding of categorical variables was used. PCA was applied to the full dataset containing data from the four sources. The first three components were used for the analysis. Figure 4.9(a) shows dataset projection on these three first components, where the source of each individual is identified. It can be observed that there is a clear dissimilarity on the distributions of each source. The variability metrics were calculated on these distributions. Figure 4.9(b) shows a 2-dimensional simplicial projection of the 3-simplex obtained with the method, which yielded the variability metrics shown in Table 4.4. The observed dissimilarity among the sources is reflected on the metrics. The 2-dimensional sphere in Figure 4.9(b) represents the upper variability bound defined by the 1R-simplex where all the pairwise dissimilarities are maximum—in such situation all points would be located in the sphere. Thus, the obtained simplex and metrics reflect a large variability among all sources, without a clear cluster of data sources defining an approximate centroid of the problem. The most outlying source corresponds to the Switzerland sub-dataset. That may be due to the data quality problems present in the dataset, such as the apparently wrong codification of missing values, the low number of individuals after the cleansing procedure, as well as the difference in the target variable.



(a) The UCI Heart Disease dataset on its three first PCA components. Data sources are identified.

(b) 2-simplex plot of variability.

Figure 4.9: Visualizations of multivariate variability on the UCI Heart Disease dataset.

		SPO ($\mathbb{O}$)			
Variable	GPD (Ω)	Cleveland	Hungarian	Switzerland	V.A.
Three first PCA components	.5840	.4753	.4647	.5195	.4477

Table 4.4: Results of multivariate evaluation on the UCI Heart Disease dataset

4.6 Discussion

4.6.1 Significance

The common methods to assess the variability of multi-source biomedical data are generally suited to univariate measurements, and most take parametric or homoscedasticity assumptions on them. The evaluation results of the variability metrics developed in this work show that these metrics are a robust alternative to classical methods on multi-type, multi-modal and multivariate data, or a complementary tool when classical assumptions are met.

The GPD metric theoretically aims to increase as the global pairwise dissimilarity among the PDFs of data sources increases. That was validated by the evaluation results. Thus, the purpose to measure the degree of variability of multi-source data is accomplished. This is analogous to classical methods, but with the advantage of being suited to multi-type, multi-modal and multivariate data. Additionally, it has been shown that the GPD keeps stable as the sample size decreases in comparison with the p-values of classical statistical methods such as ANOVA Figure 4.5.

The SPO metric provides additional information about the outlyigness of each data source with respect to a latent central tendency of all the sources' distributions. To our knowledge such information is not provided by any classical test. On numerical data, ANOVA provides the sum-of-squares measurement as a measurement of the variability between groups. That is conceptually equivalent to the intermediate PDF dissimilarity matrix obtained during multi-source variability calculus. The PDF dissimilarity matrix, however, is bounded and suited to the aforementioned features of data distributions.

Regarding to data quality, Weiskopf and Weng (2013) identified some methods to measure the *concordance* of datasets based on comparisons with gold standard equivalent repositories. The variability metrics permit measuring such degree of dataset concordance without requiring an additional gold standard dataset. Hence, the GPD metric provides the degree of concordance among datasets, while the SPO metric provides the degree of concordance of specific datasets with respect to a latent reference to all the datasets. Hence, the GPD and SPO can be defined as a composite measurement method of a multi-source variability data quality dimension. The multi-source variability can therefore be assessed under data quality assurance protocols.

One of the most practical use cases where the proposed methods can be used is the initial data understanding and data preparation stages of multi-source biobanks based research. It includes data mining or clinical trials. The GPD metric can be used to find global dissimilarities among data sources' PDFs. Large values could be caused by a low overall probabilistic concordance, or by outlying specific sources, due to possible centre or user biases. Such source outlyingness would be measured by the SPO metric. Researchers could decide to remove anomalous sources from their study or take the appropriate decisions to correct possible biases. As an example, in the development of predictive models outlying sources may reduce the global effectiveness and generalisation of models. Researchers may even consider detected variabilities as an outcome of their studies. In addition, the multi-source variability plot may help to visually identify patterns among a large number of sources, with the possibility to use the intermediate PDF dissimilarity matrix as the input of subgroup discovery algorithms such as hierarchical clustering.

4.6.2 Limitations

Using the multi-source variability metrics may require some attention under some situations, as well as in most actual data mining methods. Results showed that metrics are scalable to the number of variables. This is true according to the theoretical definition of metrics. However, in practice, the curse of dimensionality may affect to the metrics. Hence, as the number of variables increases, the probabilistic space becomes sparser. Specifically, the sparsity of a low number of data points—i.e., individuals—across the probabilistic space may cause the PDF estimations to be inaccurate—e.g., sparse, unsmoothed or 'peaky' PDFs—, leading to anomalous PDF distances. Such a variance of PDF distance estimators related to dimensionality has been discussed in other studies (Carvalho et al, 2013).

Nevertheless, as in most data mining tasks, the curse of dimensionality can be relaxed using proper dimensionality reduction methods or selecting a subset of appropriate study variables. In this work, PCA was used in the multivariate evaluation experiment. However, other non-linear methods or methods with a more intelligent treatment of categorical variables may be more suitable with multi-modal or categorical data. E.g., if distances among categories can be specified, the ISOMAP algorithm could be used to generate a dimensionality reduced manifold conserving distances between data points.

On the other hand, even when no dimensionality reduction is required, the PDF estimation method may also imply some variance on the PDF distances and, thus, to the variability metrics. The estimation of categorical histograms is straightforward. However, numerical data can be estimated using both histograms or other smoothing methods such as KDE, which may require tuning specific parameters such as the bin size (in the case of histograms) or kernel bandwith (in the case of KDE). As a consequence, an inadequate parametrization may lead to inaccurate PDFs. With the purpose to accurately estimate PDFs, parameters can be selected manually, where the optimum values are selected by a user, or automatically, using different methods to select them (Silverman, 1986; Shimazaki and Shinomoto, 2007). In this work, the KDE bandwidth was selected using the latter approach, simulating a totally automatic multi-source variability assessment. The automatic method provided reliable estimations. However, the use of other method or some manual adjustments on the kernel bandwidths may have provided slightly different results. Nevertheless, in the proposed method to obtain the variability metrics, the PDF estimation step is flexible to the use of different estimation methods suited to specific purposes or based on semantic knowledge about the problem.

Other aspect avoided in this work but which may be present on real multi-source biomedical data is the patient overlap. Weber (2013) showed that the patient overlap among different sources may limit the effectiveness of tools oriented to multi-site datasets. Thus, if it is to happen, it should be considered before applying any method. However, if the number of individuals is sufficiently high in comparison with those overlapping patients, that problem may be of little significance.

4.6.3 Future work

Some of the classical methods, such as ANOVA or χ^2 tests, have associated p-values indicating the statistical significance on the difference between the univariate measurements. They allow taking decisions based on the rejection of a null hypothesis. The variability metrics do not currently provide such a p-value, hence, its interpretation aimed to decision making may require further understanding. The GPD can be considered a estimator equivalent to the notion of normalized standard deviation of PDFs. As a descriptive estimator, further work can be carried out to characterize its measurements on different contexts and problems. First, the GPD behaviour can be characterized according to different changes on different types of distributions, as described in Chapter 3. Second, the GPD outcomes can be associated to evaluation indicators of different target problems combining multi-source data. As an example, it may help understanding which GPD thresholds are sufficient to maintain acceptable error bounds in predictive modelling combining multi-centre data. Also related to the characterization of the method, we will study the relationship between the latent central distribution provided the simplex centroid with a naïve global distribution obtained by pooling the data from all the multiple sources. This will suppose a helpful exercise to evaluate the improvements of the SPO metric respect to the distance to that naïve distribution, given that the latter will be weighted by each source sample sizes, in addition to be less informative (as the individual source features are averaged) than the centroid-latent one. Regarding to the SPO, as shown in Figure 4.2 (b), the metric appears to be convex with respect to the number of sources for a fixed distance among them, a property that can be proved in future work. On the other hand, it is also left for future work studying the possibility to provide confidence intervals on the variability metrics.

Nowadays many biomedical studies still count with low sample sizes, what may lead to the aforementioned limitations, specially in high dimensions. Hence, further work should be carried out with the purpose to characterize this effect to obtain possible calibrations or error bounds for the metrics. Additionally, such work may be combined with the study of the proper dimensionality reduction methods suited to the analysed data.

It may also be noted that as the Jensen-Shannon distance was used in this work as PDF distance for its symmetry, smoothness and bounds, that distance is at a small constant to the Hellinger distance (Jayram, 2009; Sáez et al, 2013b). Hence, each of them may be used interchangeably for the proposed metrics. Further studies may identify specific features for their selection.

Other interesting capabilities of the method emerge as future work aimed to the data preparation procedures. The method can be used to assess the variability of other data quality features such as missing data. The GPD and SPO metrics represent additional features of the dataset which may improve the development of models or hypotheses on multi-source data. In an environment with a large number of sources, such a large set of hospitals in a country, or a large number of users in a hospital, the simplicial projection can be used to obtain a clustering of these sources, as well as to provide 2D or 3D visualizations of the source dissimilarities. Hence, further visual analytics methods for data source multi-source variability will be studied to provide more informative visualizations (e.g., considering sample sizes or other source features) and interactive control panels. Finally, measuring the variability metrics through a set of temporal batches can provide a temporal monitoring of the inter-source variability as well as help to detect and monitor source biases.

Further discussions can be made deriving the application of the developed variability metrics to other purposes. Data source variability, as studied in this work, can be classified as a representation learning problem. Representation learning (Bengio et al, 2013) aims to find latent prior knowledge, namely ‘priors’, about data to facilitate the data understanding and model development on data mining problems. Hence, the GPD or SPO metrics may be used to represent such a prior knowledge of data. For instance, in a multi-source dataset each source outlyingness can be included as an additional variable to compensate possible dissimilarities on sources when developing data models. Similarly, the metalearning field of study (Brazdil, 2009) aims to find metaknowledge about models or data to guide the search of the most appropriate model for a specific problem. Thus, the variability metrics could be used to characterize particular datasets, where their effectiveness as a metaknowledge feature to choose apropriate models could be studied.

Finally, the use of the SPO as a metric to track outlyingness in temporal batches of data could also be studied. However, although it could indeed measure the degree of difference of a temporal batch with respect to others, the approach would miss the temporal relationship among those time batches—each would be treated as an independent data distribution, while they are not—, where other methods such as those proposed in the next Chapter, or a further adaptation of the SPO, would be more adequate.

4.7 Conclusions

When multi-source data samples are expected to represent the same, or a similar population, variabilities among the sources’ PDFs may hinder any data exploitation or research processes with such data. This work constructs metrics for assessing such variabilities. As an objective, the metrics should be robust to multi-type, multi-modal and multi-dimensional data as well as bounded and comparable among domains. The here developed method based on simplicial projections from PDF distances have demonstrated capabilities to accomplish these hypothesis, providing metrics for measuring the global probabilistic deviation of data, the source probabilistic outlyingness of each data source, and a interpretable variability plot visualization of the inter-source variability. The metrics can be used as a complementary or alternative method to classical univariate statistical tests, with the advantages of being independent to the type of variable, dealing with multi-modal distributions, and providing additional visualizations. Additionally, the GPD metric, Ω, stands as an estimator equivalent to the notion of the normalized standard deviation of a set of PDFs, a concept that may be used in several different purposes.

In practice, the multi-source variability metrics can be used as part of data quality assurance protocols or audit processes. The GPD and SPO metrics conform a multi-source variability data quality dimension to assess the multi-source probabilistic concordance of data, and without the need of a gold standard reference dataset. Hence, the variability metrics may help assuring the quality of—increasingly larger—biobanks-based research studies involved with multi-center, multi-machine or multi-user data.

Chapter 5

Probabilistic change detection and visualization methods for the assessment of temporal variability of data repositories

Knowledge discovery from biomedical data can be applied to on-line, data-stream analyses, or to retrospective, timestamped, off-line datasets. In both cases, variability in the processes that generate data or in their quality features through time may hinder either the knowledge discovery process or the generalization of past knowledge. This chapter establishes the temporal variability as a data quality dimension and proposes new methods for its assessment based on a probabilistic framework. Concretely, methods are proposed for (1) monitoring changes, and (2) characterizing changes, trends and detecting temporal subgroups. First, a probabilistic change detection algorithm is proposed based on the Statistical Process Control of the posterior Beta distribution of the Jensen-Shannon distance, with a memoryless forgetting mechanism. This algorithm (PDF-SPC) classifies the degree of current change in three states: In-Control, Warning, and Out-of-Control. Second, a novel method is proposed to visualize and characterize the temporal changes of data based on the projection of a non-parametric information geometric statistical manifold of time windows. This projection facilitates the exploration of temporal trends using the proposed IGT plot and, by means of unsupervised learning methods, discovering conceptually-related temporal subgroups. Methods are evaluated using real and simulated data based on the United States (US) National Hospital Discharge Survey (NHDS) dataset.

The contents of this chapter were published in the journal publication by Sáez et al (2015)—thesis contribution P4. The developed methods are included in the software contributions S1 and S3.

5.1 Introduction

Knowledge discovery on biomedical data is generally performed over healthcare repositories or research biobanks. Either when the research repositories are generated from routine clinical data or when they are specifically designed for a research purpose, it is well accepted that the efficiency of the research processes and the reliability on their information and results are improved if the repository has been assessed for data quality (Cruz-Correia et al, 2010; Weiskopf and Weng, 2013).

The data quality research has gained attention since the work by Wang and Strong (1996). Following their approach, many studies have been developed to define what characteristics of data are related to its quality, generally known as data quality dimensions. Additionally, the increasing establishment of electronic health records (EHR) and the increase of available data is wide-spreading the necessity of biomedical data quality assessment procedures for maintaining high-quality, curated biomedical information repositories (Weiskopf and Weng, 2013).

Time is a factor that has been studied in relation to the biomedical data quality in some works. Recently, Weiskopf and Weng (2013) performed a systematic review on methods and dimensions of biomedical data quality assessment. From a resultant pool of 95 articles, only four were related to the *currency* of data. According to these studies, currency refers to the degree of how up-to-date the measurements of a patient are, and it is measured based on temporal thresholds. However, Cruz-Correia et al (2010) and Sáez et al (2012b) introduced that another aspect of data quality is related to the fact that when data is collected for long periods of time, the processes that generate such data do not need to be stationary. This may be due to several reasons, such as changes in clinical protocols, environmental or seasonal effects, changes in the clinical staff, or changes in software or clinical devices. Thus, the non-stationary biological and social behaviour, as the source of biomedical data, may lead to different types of changes in data probability distribution functions (PDFs), namely gradual, abrupt or recurrent. These changes may also lead to partitions of data into subgroups of conceptually and probabilistically-related time periods, namely *temporal subgroups*. Therefore, if it is assumed that the data generating processes are stable through time, undesired and unexpected data changes may lead data to fail meeting users—i.e., data analysts—expectations, thus being considered as a lack of data quality.

This work proposes new methods for the assessment of temporal changes in biomedical data PDFs which can be used as a framework under a *temporal variability* data quality dimension. This is related to assessing the changes causing non-stationarity of data time series (Brockwell and Davis, 2009). Hence, methods are proposed to (1) monitor changes, and (2) characterize changes, trends and detect temporal subgroups. In addition, due to the heterogeneous characteristics of biomedical data (Sáez et al, 2013b), methods must be robust to different variable types, as well as to multivariate, multi-modal data. Furthermore, in order to improve the scalability of the methods, their outcomes should be provided as comparable among different domains, hence requiring bounded metrics. As a consequence, a probabilistic framework is established to support the proposed methods, comprising (1) a non-parametric synopsis of PDFs, using an incremental, memoryless non-forgetting approach (Rodrigues et al, 2010), and (2) a PDF distance measurement based on information-theoretic probabilistic distances (Csiszár, 1967; Lin, 1991), concretely in the Jensen-Shannon distance (Endres and Schindelin, 2003).

The first proposed method is a probabilistic change detection algorithm to monitor changes in non-parametric PDFs through time. It is based on the concepts of the Statistical Process Control (SPC) by Gama et al (2004), originally designed for drift detection in the performance of machine learning models. The new algorithm (PDF-SPC) is based on the monitoring of the incrementally estimated posterior Beta distribution of the Jensen-Shannon PDF distance, classifying the degree of current change in three states: In-Control, Warning, and Out-of-Control.

The second proposed method is a novel approach to visualize and characterize the temporal changes of data based on the projection of a latent, non-parametric information-geometric statistical manifold (Amari and Nagaoka, 2007) of time windows. Concretely, a dissimilarity matrix is obtained from the PDF distances among the different time windows, where multi-dimensional scaling is used afterwards to project the temporal statistical manifold (or a dimensionally reduced version). Hence, being the PDFs of time windows projected in a geometric space, this permits visualizing and characterizing the temporal changes that occur in data, as well as to apply unsupervised learning methods, such as clustering, to obtain conceptually-related subgroups of temporal windows.

The interpretation of the results provided by the proposed methods is facilitated by visual methods, namely PDF-SPC control charts and information-geometric temporal plots (IGT plots) of the statistical manifolds. Also, dendrograms and dissimilarity heat maps can be used as complementary visualizations. Additionally, as a by-product of the probabilistic framework, the continuous estimation of PDFs leads to *probability mass temporal maps* which, similarly to spectrograms, help understanding the temporal changes of probability distributions.

The rest of the chapter is organized as follows. Section 5.2 describes the required background. Section 5.3 describes the probabilistic framework and the two proposed methods. Section 5.4 describes the National Hospital Discharge Survey (NHDS) dataset used in the evaluation. Section 5.5 describes the evaluation and its results. Section 5.6 discusses the study, and compares it with state-of-the-art related work. Finally, Section 5.7 provides the conclusions of the chapter.

5.2 Background

Biomedical data are generally gathered in two ways for its analysis: *on-line* and *off-line*. The classical method for accessing research data is as an off-line dataset, e.g., a comma-separated values file or a small relational data base. However, the continuous increase on the amounts of available clinical data is changing the tendency to on-line methods, where data is analysed through the continuous observation of batches, generally aiming to optimise processing and storage resources (Gama and Gaber, 2007; Rodrigues and Correia, 2013). On the other hand, when the purpose is to monitor biomedical indicators in real time, the on-line analysis is straightforward. This work aims to apply to both scenarios, thus, providing users feedback about changes on their off-line dataset or during on-line processes.

This section describes some previous theoretical background which is required for the new methods proposed in this work. Concretely, this background is divided in two main topics: the probabilistic framework to compare biomedical data distributions and the change detection methods.

5.2.1 Probabilistic distances on biomedical data distributions

Biomedical data show heterogeneous conditions. They are generally based on multimodal distributions—i.e., various inherent generative functions, such as a mixture of affected and unaffected patients. Studies may be uni- or multivariate, and may be composed of different types of variables—i.e., continuous, discrete ordinal and non ordinal, or mixed. Under these conditions, comparing different data samples or batches through time based on classical statistics may not be enough representative, or even not valid. Additionally, in order to measure the magnitude of changes it is interesting to provide a metric for such comparisons which, ideally, should be bounded to facilitate its comparability on different domains.

In Chapter 3 we studied the behaviour of different PDF dissimilarity metrics with respect to these conditions. The results of such study are summarized in Table 3.1. The results showed that the aforementioned data features may complicate the application of classical statistical or data analysis methods for the assessment of differences among data samples. Specifically, the results confirmed that classical statistical tests may have difficulties on multi-modal data, or may not be suitable at all on multivariate or multi-type data. Information-theoretic distances, including the Jeffrey and Jensen-Shannon distances, and the EMD resulted the most suitable distances to all conditions. Information-theoretic are distances which derive from the Shannon's entropy theory, while EMD derives from the digital imaging field as the optimal minimum cost of transforming one histogram into another. Then, information-theoretic distances permit constructing over the theory of a probabilistic framework.

Focusing on the information-theoretic distances, the Jeffrey distance is a symmetrized, metric version of the Kullback-Leibler divergence. However, it is not bounded and, as we showed in Chapter 3, when the probability mass in any region of the support in any of the compared PDFs tends to zero, the metric tends to infinite. In contrast, the Jensen-Shannon distance (JSD), square root of the Jensen-Shannon divergence, is a metric bounded between zero and one, and it was smoothly convergent to one on that situation. As a consequence, in this work the JSD was selected as the distance between PDFs.

5.2.2 Change detection

Change detection methods have been widely studied in data streams, specially when data are generated as a continuous flow and limited processing or storage resources are available (Gama, 2010). Change detection aim at identifying changes on sufficient statistics of the sample measured through time (Basseville and Nikiforov, 1993; Gama and Gaber, 2007). The selection of the change detection method and the corresponding sufficient statistic depend on the purpose, and generally follow two approaches: (1) monitoring data distributions, such as the evolution of the average; or (2) monitoring the evolution of performance indicators, such as the fitness of data mining models or patterns (Klinkenberg and Renz, 1998).

Changes can be classified according to their causes and to their behaviour, e.g., their rate of change. Regarding to the causes, changes can occur due to modifications in the context of data acquisition, e.g., changes in clinical protocols. On the other hand, related to their behaviour, changes may be characterized as 1) gradual, 2) abrupt and 3) recurrent. In the literature, gradual changes are also known as *concept drifts*, while abrupt as *concept shift*. Abrupt changes do not necessarily imply changes with a large magnitude. In fact, the early Warning of small changes may be of crucial importance to prevent larger problems caused by the accumulation of such small changes (Basseville and Nikiforov, 1993). The proper change detection methods will depend on these requirements.

In this study, the focus is to detect and characterize changes in the PDF of data. This is usually based on monitoring temporal windows of the current PDF with respect to a reference window. This involves three related aspects: (1) the type of window scheme, (2) the synopsis of the windowed data into a sufficient statistic, and (3) the change detection method on the sufficient statistic.

Time windows: Time windows schemes define the characteristics of the temporal period which data is synopsed—i.e., aggregated—to be monitored. The simplest approach is to use sliding windows of fixed size (Mitchell et al, 1994; Gama and Gaber, 2007). Thus, for a window size of w observations, when the individual i is observed, the $i-w$ is forgotten. This approach is useful on sensor data, which are expected to arrive in a continuous stream. However, biomedical data do not necessarily have a constant flow, e.g., the number of patient discharges presents large variations during the day, among the days of the week, or have a seasonal effect. This, in addition to the social organization of time, may lead to an inaccurate statistical sampling. Hence, a solution comes by using sliding windows within temporal semantic landmarks (Gehrke et al, 2001), i.e., aggregating daily, weekly, or monthly data, independently of the number of individuals within each semantic block. These approaches use a catastrophic-forget, i.e., the outside-window information is ignored. However, as concepts may evolve smoothly, old data may still be important (Gama et al, 2004). In tilted windows, current information is an aggregation at increasing levels of granularity from past to current data (Han et al, 2012). Thus, old data are still used but latter examples are given more importance. Other approach to synopse data without forgetting is using weighted sliding windows. Hence, each observation is weighted according to its age, getting older data less weight. Due to the fact that the amount of memory is limited, specially in ubiquitous streams scenarios, weighted sliding windows weight data individuals within a window. Thus, there is still a minor outside-window forgetting. In order to overcome this issue, Rodrigues et al (2010) proposed the incremental memoryless fading windows. It uses all previous data in an incremental manner, i.e., only the last observation is maintained in memory, approximating weighted windows within specific error bounds.

Synopsis: Synopsis methods aim to aggregate or summarize data within a window as the basis for the sufficient statistic to be monitored for changes. Simplest methods may just calculate the window central tendency—e.g., a weighted average according to weighted windows schemes—and dispersion. In scenarios where the Gaussian behaviour is not the default, other methods such as histograms or wavelets (Chakrabarti et al, 2001) may result more suitable. Thus, frequency histograms or wavelet coefficients are calculated on the window data as a compact aggregation of its information. The synopsed sufficient statistic of a current window i can be calculated, as previously mentioned, according to weighted past information. Hence, memoryless fading windows provide α-fading sufficient statistics considering all previous data points. The general form of an α-fading statistic $\Upsilon_\alpha(i)$ over a sequence of observations $\{\upsilon_i\}$ is

$$\Upsilon_\alpha(i) = \begin{cases} \upsilon_1, & \text{i} = 1, \\ \upsilon_i + \alpha \cdot \Upsilon_\alpha(i-1), & \text{i} > 1, \end{cases} \tag{5.1}$$

with $0 < \alpha < 1$. Hence, $\Upsilon_\alpha(i)$ is the α-fading statistic obtained from the synopsis of data in the landmarked window i.

Detection: Change detection methods have been proposed depending on the type and purpose of the analysed data (Sebastião and Gama, 2009). The Page-Hinkley Test (Mouss et al, 2004) is one of the most referred when the monitored data is assumed to show a Gaussian behaviour—e.g., in industrial processes. Data streams do not necessarily need to follow a Gaussian distribution. To deal with this, Kifer et al (2004) proposed a non-parametric change detection method based on a relaxation of the total variation distance between PDFs. This is important when change detection is to be applied not to a single data stream, but to a non-parametric probability distribution, and it is specially a challenge when monitoring multivariate sets of data with multiple types of variables simultaneously. On the other hand, with foundations on the Statistical Quality Control by Shewhart and Deming (1939), Gama et al (2004) proposed a Statistical Process Control method to detect changes in the performance indicators of machine learning models—i.e., the classification error-rate. Their SPC defines three possible states for the process: In-Control, Warning and Out-of-Control. The state is selected according to the confidence interval of the current error-rate to be generated from the original distribution. Thus, an Out-of-Control state is associated to a concept drift, leading to the re-learn of a new classification model with the observations since the last Warning state—as a meaningful reference of the beginning of the new concept.

5.3 Proposed methods

This section describes the proposed methods for the assessment of the temporal variability DQ dimension. The proposed methods are based on a common probabilistic framework defined by the measurement of the distance between the PDF of different temporal windows. This framework is described first in this section. Then, the new methods for change monitoring and for the characterization and subgroup discovery are described.

5.3.1 Probabilistic framework

The framework defines the methods to (1) estimate the PDF of the data within a window, and (2) measure the PDF distance between two windows.

In terms of change detection, the method to estimate the window PDF can be defined according to a time window scheme and synopsis method. A prior consideration is that the social organization of time is reflected in temporal biomedical data. Thus, depending on the hour, weekday, week, month or year there will always be an implicit biased behaviour. As a consequence, the use of such a temporal landmarked windows (with a granularity according to the characteristics of the study) is recommended for a proper sampling. Hence, sufficient statistics will aggregate the data within such windows.

On the other hand, in Section 5.2.2 it was defined that the flow at which biomedical data is generated is not generally constant—i.e., the number of individuals per time period. This may depend on the aforementioned social organization, but also on other contextual factors. Therefore, the data samples in different landmarked windows may not be enough representative, and may lead to inaccurate sufficient statistics. In order to overcome that issue, the landmarked windows are combined with a memoryless fading windows scheme. The initial landmarked window and the fading window are used in different tasks. While the former contains the data points which are synopsed to obtain the sufficient statistic, the later contains the set of sufficient statistics which are gradually weighted. In addition to the computational advantages of memoryless fading windows, as an approximation to weighted windows they contribute to the non-forgetting of past data, which is important for the tracking of gradual changes.

A requirement for the temporal variability methods is that they must be robust to the heterogeneous conditions of biomedical data. Hence, synopsis methods should capture such information for further analyses. With such a purpose, histograms stand as a proper method as they can be obtained for continuous, discrete, and even for mixed types problems, as well to multivariate data.

On discrete variables, histograms may exactly correspond to their PDF, where each bin contains the probability mass associated to a value on the distribution support. However, on continuous distributions histograms must be defined according to a set of non-overlapping intervals, leading to a discrete number of bins approximating the original continuous PDF. Different techniques exist to obtain the proper number of bins on continuous data (Guha et al, 2004; Shimazaki and Shinomoto, 2010). Additionally, when the problem is purely continuous, KDE methods (Parzen, 1962; Bowman and Azzalini, 1997) can be used to obtain a generative and smoothed PDF.

As a consequence, each window PDF, further on P_i, will be approximated as an α-fading averaged histogram where the probability mass of each bin is

$$H_{b,\alpha}(i) = \frac{S_{b,\alpha}(i)}{N_{b,\alpha}(i)}, \tag{5.2}$$

where, following Equation 5.1, $S_{b,\alpha}(i)$ is the α-fading sum of the raw probability mass of bin b at window i, defined as

$$S_{b,\alpha}(i) = \begin{cases} p_b(1), & \text{i} = 1, \\ p_b(i) + \alpha \cdot S_{b,\alpha}(i-1), & \text{i} > 1, \end{cases} \tag{5.3}$$

where $p_b(i)$ is the probability mass of bin b at window i. Besides, $N_{b,\alpha}(i)$ is the corresponding α-fading increment (i.e., the α-fading account of averaged bins), and is defined as

$$N_{b,\alpha}(i) = \begin{cases} 1, & \text{i} = 1, \\ 1 + \alpha \cdot N_{b,\alpha}(i-1), & \text{i} > 1. \end{cases} \tag{5.4}$$

The memoryless approximation of the α-fading averaged histogram is not error free in comparison to a weighted approximation. It is proved that the error can be bound within a confidence interval of $\pm 2\epsilon R$ setting $\alpha = \epsilon^{\frac{1}{w}}$, where $R = 1$ is the variable range—as a probability mass—, and w corresponds to the window size to approximate (Rodrigues et al, 2010).

On the other hand, the framework establishes a method for the measurement of the distance between the PDFs of two windows. Such method should be 1) robust to multivariate, multi-type and multi-modal data, 2) bounded and 3) smoothly convergent with near-0 probability bins. As discussed in Section 5.2.1, and according to the results of Chapter 3, a method that fulfils these properties is the Jensen-Shannon distance. Hence, the distance between the PDFs of two windows, P_i and P_j is

$$d(P_i, P_j) = JSD(P_i||P_j), \tag{5.5}$$

where $JSD(P_i||P_j)$, is the Jensen-Shannon distance in equation 2.36. The used Jensen-Shannon distance is based on the Kullback-Leibler divergence (Equation 2.29) which, considering the histogram approximation of the PDFs, will be calculated as:

$$KL(P||Q) = \sum_b \log_2 \left(\frac{P_b}{Q_b} \right) P_b, \tag{5.6}$$

where P_b and Q_b are the approximated probability mass at bin b. We recall that using the base 2 logarithm to calculate the Kullback-Leibler divergence, the Jensen-Shannon distance is bounded between zero and one.

5.3.2 Change monitoring

With the purpose of monitoring changes as part of the temporal variability data quality assessment, a new change detection algorithm is proposed. The degree of change between the PDFs of two time windows is given by their Jensen-Shannon distance. The JSD is $[0, 1]$-bounded and always positive. Thus, in a stable process, i.e., where the data distribution under study only varies over time within some small noise, monitoring the JSD between the PDFs of the current window and a reference past window will provide a stable signal close to zero. Then, the objective would be monitoring a sufficient statistic associated to the data variability, i.e, a sufficient statistic of the distribution of the Jensen-Shannon distances. The proposed change detection and monitoring method is based on the concepts of SPC by Gama et al (2004)—originally aimed to monitoring the error rate of predictive models—to monitor the data variability based on the Beta distribution of the JSD.

Suppose a sequence of PDF estimations $\{P_i\}$. Using the first element as a reference, $P_{ref} = P_1$, the JSD of further elements $P_2, ..., P_i$ with respect to the former provides a sequence of distances $\{d_i\}$. In a stable process, d will approximately be distributed around a central tendency measure close to 0 associated to a latent noise. More strictly, as the JSD is $[0, 1]$-bounded, d can be defined by a Beta random variable. Hence, in a stable process, after a transitory state, the mean value μ of the $Beta(\alpha, \beta)$ distribution B given by $\{d_i\}$ will remain stable. Additionally, an upper confidence interval u^z for B is given by the inverse cumulative distribution function $iCDF(.5 + z/2)$, with $0 < z < 1$—e.g., for an upper confidence interval at 95% then $z = .95$.

The proposed PDF-SPC method manages three registers during the monitoring, u^{z1}_{min}, u^{z2}_{min} and u^{z3}_{min}, with $z_1 < z_2 < z_3$. For each new distance d_i, which updates the Beta distribution B, if the new u^{z1}_i is lower than u^{z1}_{min}, the three registers are updated based on B_i. Hence, the values of z_1, z_2 and z_3 depend on the desired confidence levels. In this work, we have established those confidence levels based on the three-sigma rule, therefore, the upper confidence intervals are set to $u^{.68}_{min}$, $u^{.95}_{min}$ and $u^{.997}_{min}$. This decision is based on widely adopted confidence intervals in statistical methods, however, the selection of confidence levels may be adapted to specific domains of use, or even calibrated according to a desired response.

Given a new distance d_i, three possible states are defined for the process:

- In-Control: while $u^{z1}_i < u^{z2}_{min}$. The monitored PDF is temporary stable.
- Warning: while $u^{z1}_i \geq u^{z2}_{min} \wedge u^{z1}_i < u^{z3}_{min}$. The monitored PDF is changing but without reaching an action level. Its causes may be noise or a gradual change. Hence, an effective change should be confirmed based on further data.
- Out-of-Control: whenever $u^{z1}_i \geq u^{z3}_{min}$. The current PDF has reached a significantly higher distance from the past reference. The current B_i is different from the reference with a probability of z_3.

Reaching the Out-of-Control state means that a new concept is established. As a consequence, in order to continue the change monitoring, the PDF-SPC algorithm (Algorithm 1) will replace the reference PDF with the current concept. Hence, if the Out-of-Control state is reached after P_j is observed, then $P_{ref} = P_j$.

As well as the estimation of PDFs is based on a α-fading incremental approach, with the purpose to avoid storing in memory all the observations of d_i, the distribution B is updated using an incremental approach. Hence, the estimation of the parameters of B, $\hat{\alpha}$ and $\hat{\beta}$ is based on the Maximum Likelihood Estimation (Hahn and Shapiro, 1968) where the initialization of the parameters (Equations 5.7 and 5.8) was modified to use a recursive estimation of the sample geometric mean, $\hat{G}$ (Equation 5.9).

$$\hat{\alpha} = \tfrac{1}{2} + \frac{\hat{G}(d_i)}{2(1 - \hat{G}(d_i) - \hat{G}(1 - d_i))} \tag{5.7}$$

$$\hat{\beta} = \tfrac{1}{2} + \frac{\hat{G}(1-d_i)}{2(1-\hat{G}(d_i)-\hat{G}(1-d_i))} \tag{5.8}$$

$$\hat{G}(x_i) = \left(\left(\hat{G}(x_{i-1})\right)^{i-1} x_i\right)^{1/i} \tag{5.9}$$

```
input: $P_{ref}$, current reference PDF
Sequence of PDFs: $\{P_i\}$
begin
    Let $P_i$ be the current PDF
    Let $d_i = JSD(P_i||P_{ref})$
    Let $B$ be a $Beta(\alpha, \beta)$ distribution
    Re-estimate $B$ with $d_i$
    if $u_i^{z1} < u_{min}^{z1}$ then
        $u_{min}^{z1} = u_i^{z1}$
        $u_{min}^{z2} = u_i^{z2}$
        $u_{min}^{z3} = u_i^{z3}$
    end
    if $u_i^{z1} < u_{min}^{z2}$ then
        // In-Control
        $Warning? \leftarrow False$
    else
        if $u_i^{z1} < u_{min}^{z3}$ then
            // Warning Zone
            if $NOTWarning?$ then
                $Warning? \leftarrow True$
            else
                nothing
            end
        else
            // Out-of-Control
            $P_{ref} = P_i$
            $Warning? \leftarrow False$
            Re-start $B$, $u_{min}^{z1}$, $u_{min}^{z2}$, $u_{min}^{z3}$
        end
    end
end
```

Algorithm 1: The PDF-SPC change monitoring algorithm

The PDF-SPC permits identifying timestamps related to concept changes, i.e., whenever Warning and Out-of-Control states are reached. As a possible initial indicator of further larger changes (Basseville and Nikiforov, 1993), that information is specially useful to rapidly react to, or even to predict, changes. On the other hand, Widmer and Kubat (1996) suggested that two concepts may coexist before a change is achieved. The Warning state is fired when there is a suspect for a change, which may be confirmed once there is enough evidence by the Out-of-Control state. Hence, the temporal distance between a Warning an Out-of-Control states may be an indicator of such period of coexistence of concepts and, thus, of the rate of change. However, other descriptive information to characterize the behaviour of changes may be missed, e.g., whether concepts can be grouped into meaningful, possibly recurrent, groups. A promising novel method to deal with this problems is described next.

5.3.3 Characterization and temporal subgroup discovery

With the purpose to characterize the behaviour of changes and facilitate the discovery of temporal subgroups, a novel method is proposed. As the PDF-SPC monitors the degree of changes, this new method aims to describe them, facilitating their characterization, e.g., into gradual, abrupt or recurrent, and analysing the evolution of data inherent concepts.

According to the probabilistic framework, each time window can be seen as an individual characterized by its PDF estimation. The Information Geometry field states that probability distributions lie on a Riemannian manifold whose inner product is defined by the Fisher Information Metric of a specific family of probability distributions (Amari and Nagaoka, 2007). The geodesic distances between the points associated to PDFs are approximated by their PDF divergences, such as the Jensen-Shannon. Hence, the JSDs among each pair of PDFs can be used to approximate a non-parametric—i.e., family-independent—statistical manifold where the temporal PDF estimations lie and, as a consequence, allow the discovery of related trends and subgroups. In addition, due to the JSD bounds, the maximum possible distance among any pair of PDF points is one. That means that the approximated statistical manifold is bounded by a hyper-ball of diameter one. Hence, the studied PDFs will lie on space comparable among different problems, as it will be known that: 1) equal PDFs will co-locate and 2) completely separable PDFs will be located at the hyper-ball surface—i.e., at a distance of one.

Suppose a sequence of PDF estimations $\{P_i\}$, with $1 < i < n$. The $\binom{n}{2}$ pairwise distances $d(P_i, P_j)$ define a n-by-n symmetric dissimilarity matrix $Y = (y_{11}, ..., y_{nn}), y_{ij} : d(P_i, P_j)$. Hence, Y can be used as the input of a compatible[d] clustering method, such as a complete linkage hierarchical clustering, which will provide a set of groups G_k, each related to a data inherent temporal concept.

The approximated statistical manifold provides information about the layout of PDFs in such a latent space, e.g., to discover conceptual subgroups. However, much more information can be taken considering that there is an implicit temporal order among such PDF points. While the distances among subgroups indicate the concept dissimilarity, the layout of the temporal order among their points provides information about how concepts evolved through time. Hence, a temporal continuity through the points of a subgroup, e.g., along the vector defining its largest variance, is an indicator of a gradual change. On the other hand, a temporal alternation among different subgroups every certain time period may be an indicator of recurrent abrupt

[d] Note that Jensen-Shannon distances are not euclidean, hence, compatible clustering methods or euclidean transformations should be used.

changes among probabilistically distinguished concepts. Similarly, a temporal fluctuation through a direction within a subgroup, e.g., along one of its variance vectors, may be an indicator of a recurrent gradual change among closer, probabilistically-contiguous concepts.

Hence, in order to permit such analysis it is needed to translate the dissimilarity matrix Y into a set of points in a geometric space. Considering that the distances in Y are not euclidean, the use of the MDS method is suitable to obtain an embedding of the PDFs into a euclidean space—see Section 2.2.4.

Hence, given the dissimilarity matrix Y, MDS will obtain the set $P = (\mathbf{p}_{11}, ..., \mathbf{p}_{nc})$ of points for the n PDFs in a $\mathbb{R}^c$ euclidean space such that $c = n - 1$.

Therefore, based on the calculus of a dissimilarity matrix among the PDFs of time windows or batches, and a dimensionally reduced MDS projection into 2 or 3 dimensions, we facilitate the processing of such information in an Information Geometric Temporal (IGT) plot. The IGT plot stands as a powerful visual analytics tool to explore, characterize and understand changes from a probabilistic perspective. The IGT plot then consists in a temporal statistical manifold, which PDFs are lied out as points which can be labelled e.g. with their temporal index (as shown in this chapter) or with a formatted date (what will be shown in the following chapters). Illustrative examples of such visualization are shown in the next section.

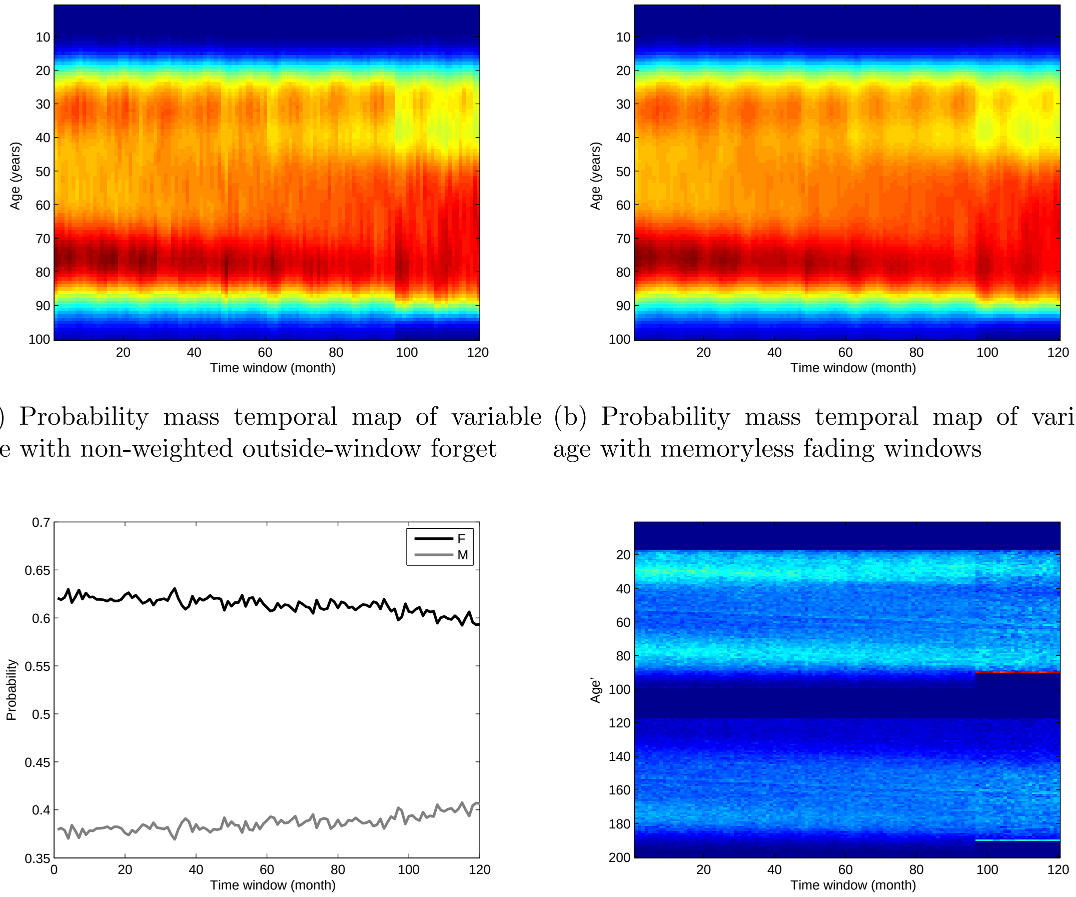
5.4 Data

This section describes the data used to evaluate the proposed methods and proposes a visualization method for monitoring PDFs.

The data used in the evaluation is the publicly available NHDS dataset (NHDS, 2014). Using only adult patients (age $\geqslant$ 18), the dataset contains 2,509,113 hospital discharge records of approximately 1% of the US hospitals from 2000 to 2009. The minimum date granularity is the discharge month. Hence, the following experiments are based on a monthly basis aggregate landmarked windows, with a total of 120 months (the time windows will be referred further on as their month index). The NHDS dataset contains several demographic, diagnosis and discharge status information. However, for the purpose of this evaluation the age and sex variables are sufficiently representative, as it is shown next.

With the purpose to illustrate the examples a *probability mass temporal map* visualization is proposed, which results as a novel visual method for the monitoring of biomedical variables. It is based on the idea of *dense pixel* visualizations (Keim, 2000), where the range of possible values are associated to a coloured pixel according to a user-specified colormap. That method has already been used to visualize sensor monitorings (Rodrigues and Gama, 2010). In this case, the method is adapted to visualize the evolution of PDF estimations, where the domain axis identifies the temporal window and the range corresponds to the probability bins. Hence, each row of the map can be seen as a signal of the probability mass evolution for a given support value. In principle, the method is suitable to variables where there is an order in the variable support, i.e. as numerical data or discrete ordered. However, it may also be useful to visualize the joint probability of ordered and non ordered variables, using the latter to divide the range axis of the map on repeated supports of the former.

Figures 6.12(a) and 6.12(b) show probability mass temporal maps of the age variable (given in years). As a univariate numerical variable, PDFs at each window are estimated based on KDE to obtain an smoother histogram. In Figure 6.12(a), an outside-window forgetting window scheme is used. In Figure 6.12(b), the memoryless fading window scheme is used (an error of $\epsilon = 0.05$ was used with a smoothing window of 12 months). It can be seen that the non-forgetting approach of fading windows leads to a smoother temporal estimation, which may avoid undesirable noise caused by non-representative windows. Note that the Gaussian-kernel estimation of KDE causes that some probability mass from the lower tails of continuous Gaussian kernels is given to the bins under 18 years.



(a) Probability mass temporal map of variable age with non-weighted outside-window forget

(b) Probability mass temporal map of variable age with memoryless fading windows

(c) Temporal evolution of the probability of variable sex (F: female, M: male)

(d) Temporal evolution of the joint probability of variables age and sex (the 2-dimensional joint distribution was vectorized partitioning by sex as: 0-100 for female age, 101-200 for male age)

Figure 5.1: Visualizations for the monitoring of the NHDS variables

It can be observed that the age variable shows a multimodal behaviour which contains several temporal artefacts of special interest for evaluating change monitoring methods. First, there is an abrupt change in month 97. That change is documented (NHDS, 2010) as a change in the codification of the variable, where ages 91 and over were recoded to 90. Second, a gradual shift is observed as mid-age patients (age $\approx$ 55) get more probability mass through time. This has associated a decrease in the mass of younger and older patients. This change may be related to a long-term contextual change—e.g., socio-economic change or due to changes in the sampled hospitals/population—possibly associated to the increase in life expectancy in the US (National Research Council, 2011; Arias, 2014). Finally, a recurrent change is observed in young (age $\approx$ 30) patients with a periodicity of 12 months. That change is associated to the increase of births in the summer period, possibly due to the live births seasonality documented for the US (Cesario, 2002) and other countries (Wellings et al, 1999). Such effect can be observed in the marginal map for female patients (see Figure C.1).

On the other hand, Figure 5.1(c) shows the temporal evolution of the probability masses of the sex variable. There is a minor gradual change in the probabilities of male and female patients, due to the increase in the males/females ratio in the US during the period of study (Howden and Meyer, 2011).

Finally, Figure 5.1(d) shows the evolution of the joint probability of age and sex. In this case, while including a discrete non-ordered variable—namely categorical—, it is not possible to directly apply KDE. Hence the raw synopsed histogram is used instead. In addition, to facilitate the visualization, the 2-dimensional histogram was vectorized partitioning by the sex variable, hence the upper half of the map identifies the evolution of age in females, and the lower in males. As a raw, non-KDE smoothed histogram, it can be observed that, first, ages below 18 do not get mass (as only adult patients were included), and second, the aforementioned change in month 97 makes the age of 90 get the highest mass (as it includes any older patients). That large difference in probability masses causes that lower values get close colors in the map. In this case, the visualization can be improved applying a logarithm function with a tuning parameter to the array of PDFs to visualize, $\log(\mathbf{P} + z)$, which assigns larger values of the colormap to the intermediate masses.

Hence, the combination of the age and sex variables in the evaluation study accomplishes the three heterogeneous characteristics of biomedical data to which methods must be robust: age is clearly multimodal, each is of different type, and changes can be studied on their joint—i.e., multivariate—distribution.

5.5 Evaluation

In this section the proposed methods are evaluated with the real changes present in the NHDS data described in previous section, as well as with simulated changes applied on it.

5.5.1 Change monitoring

The PDF-SPC algorithm was applied first to a continuous univariate problem based on the age variable with the purpose to evaluate its behaviour with respect to the present changes. Second, it was evaluated on the sex variable, categorical, where a small gradual drift occurs. Then, it was evaluated on the multivariate and mixed-types problem based on the joint probability of age and sex. Finally, a simulated abrupt shift was introduced in the latter problem as a change in the joint probability of age and sex but not in their respective univariate estimates, with the purpose to evaluate the behaviour of the SPC algorithm on that multivariate change. The confidence levels were set to $z_1 = .68$, $z_2 = .95$ and $z_3 = .997$.

Figure 5.2 shows the results of these four evaluations. The age variable monitoring, Figure 5.2(a), shows that the three types of changes are detected. First, after the transitory state there is a continuous increase in the PDF distance with respect to the reference window, associated to the gradual movement of mass to the mid-age range. This leads to a Warning state in month 46. Second, the abrupt change in month 97 was clearly detected. Third, the recurrent change on age ≈ 30 is captured as a periodic change in the probabilistic distance to the reference, however, the selected confidence levels avoid firing any change from them.

The sex variable monitoring, Figure 5.2(b), shows the gradual switch as an increase in the monitored distances. However, as expected the magnitude of the change is much lower—note that the Jensen-Shannon distance is $[0, 1]$-bounded, hence magnitudes are comparable. The recurrent change which was easily observed in the age variable is detected in this case as well. Given the 12-month periodicity, the phase displacement with respect to the age monitoring may just be due to the selected reference window.

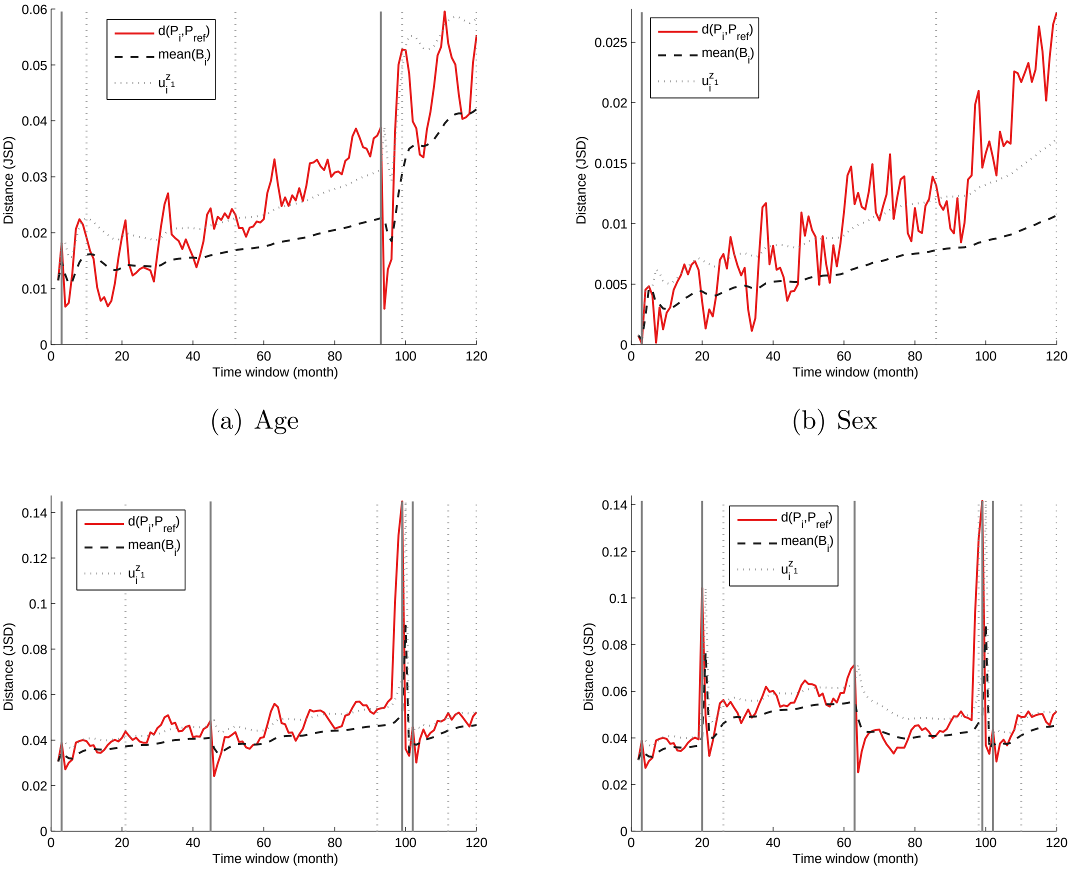
The monitoring of the joint probability of age and sex, Figure 5.2(c), also captures a gradual change as the mean distance also increases. However, maybe due to the sum of changes in both variables causes the change to be detected before. Hence, a change is fired after month 43. Additionally, the codification change in age is also detected, although a couple of iterations later.

In the last experiment, a multivariate change was introduced in month 20, maintaining the new concept until the end. Thus, the sex of n patients was switched, where n corresponds to the minimum amount of patients from any of the two sexes at each time window—males in all cases. Whilst the change is not detected univariately, the multivariate monitoring clearly detects the change in month 20.

5.5.2 Characterization and temporal subgroup discovery

The proposed methods for change characterization and temporal subgroup discovery were applied to two of the previous scenarios: in the age variable monitoring and in the monitoring of joint age and sex variables with simulated change. It is expected that characterizations and subgroups are related to the concept changes detected by the SPC method.

Figure 5.3(a) shows the 2-dimensional IGT plot associated to the statistical manifold where the temporal PDFs estimated from variable age lie. Each PDF is represented as the index corresponding to its temporal window—i.e., the month—, allowing temporal changes to be characterized. It can be observed that there are two well differentiated groups which, looking at their indices, correspond to the concepts before and after the codification change in month 97. Looking at the first subgroup, there is a linear temporal continuity through its larger variance (Arrow A). That continuity is a



(a) Age

(b) Sex

(c) {Age, Sex}

(d) {Age, Sex} with forced joint change

Figure 5.2: Control charts of the PDF-SPC evaluations. Vertical dotted lines indicate the entering into a Warning state. Vertical continuous lines indicate a change detection as an Out-of-Control state.

clear indicator of the gradual change moving probability mass to the mid-age patients. For the visual representation, a colormap has been used to assign cooler and warmer colors to winter and summer months, respectively. Hence, it can be observed that the second variance component of the first subgroup (Arrow B) is associated to the mentioned 12-month periodic change. In addition, the same change direction is shown in the second subgroup.

The apparent subgroups were confirmed with a complete linkage hierarchical clustering based on the PDFs dissimilarity matrix. Figure 5.3(b) shows a heat map of the symmetric PDFs dissimilarity matrix, where the color temperature represent a larger probabilistic distance between the PDFs P_x and P_y. The differences among the two main groups can be observed, as well as a recurrent distance increase with a 12-month periodicity. Figure 5.3(c) shows the dendrogram obtained from the clustering method, which confirms such temporal groups.

On the other hand, Figure 5.4(a) shows the IGT plot of the second scenario. In this case, it can be observed that there are three well differentiated groups. Looking at the indices, the first and second subgroups are separated by the forced multivariate change in month 20, while the second and third by the univariate change in age. In addition, there are transitory PDFs between the groups, which may be due to the smoothing produced by the fading windows approach. Figures 5.4(b) and 5.4(c), as in the previous example, confirm the discovered temporal subgroups. In this scenario, the 12-month recurrent change is hindered by the magnitude of the other changes, however, it can slightly be observed in the variance directions of the subgroups (note the separability among the seasonal colors) as well as in the dissimilarity matrix heat map. However, the change detected in month 62 does not establish a subgroup change. Nevertheless, it represents the resultant change from an accumulated gradual change on both variables whose magnitude, as it can be observed in Figure 5.2(d), is lower. The equivalent results for the age+sex scenario without the forced change are shown in Figure C.4.

5.6 Discussion

This section highlights the significant points of this work, discusses it with related work and the limitations of the proposed methods and, finally, suggests future lines of work.

5.6.1 Significance

First, the PDF-SPC algorithm has shown to accurately detect the changes present in the evaluated data according to their evidences (Section 5.4). The resultant monitoring charts provide information about the magnitude and type of changes, showing the current probabilistic distance with respect to the reference concept. Based on the Jensen-Shannon distance, the magnitude of changes is $[0 - 1]$-bounded and hence comparable among different problems, e.g., the magnitude of changes in sex (Figure 5.2(b)) is probabilistically an average of half the magnitude in age (Figure 5.2(a)). Additionally, based on an incremental approach, the method is suitable to on-line analyses with a reduced storage and computational cost.

Second, the methods for change characterization and temporal subgroup discovery based on information geometry have shown to detect temporal subgroups present on the evaluated data, as well as to help characterizing the type of changes based on the temporal tendencies of the data points associated to the PDFs of time windows. To the knowledge of the authors, this is the first study of non-parametric change detection and characterization based on information-geometric statistical manifolds, with the potential to be an important step forward. To date, most change detection methods provide information about the magnitude of changes, their classification according to the rate of change, which regions in the variables of study show a major contribution to changes, or even aim to their prediction. However, the temporal projection of a non-parametric information-geometric statistical manifold constructed from consecutive time-windows permits describing and analysing the behaviour of changes, as the evolution of a probabilistic concept through such manifold. In this study, the method has been used on one hand to construct the IGT plots, as a novel visualization tool for the exploration of temporal changes in data PDF. In other hand, the obtained PDF points have been used for unsupervised learning purposes with the purpose to find temporal subgroups. However, these are only the first steps of many further research possibilities which still remain opened based on this approach.

Analysing the two proposed methods together, the evaluation results have demonstrated the consistency between the PDF-SPC change monitoring algorithm and the information-geometric based methods for the characterization and subgroup discovery, since the change levels and detections in the PDF-SPC monitoring are associated to the obtained temporal characterization and subgroups, including the three types of changes: gradual, abrupt and recurrent. In addition, both methods result suitable to the heterogeneous biomedical data conditions posed as requirements. The use of the probabilistic distances approach permits measuring changes in multi-modal distributions, as previously demonstrated by (Sáez et al, 2013b). This, in combination with synopsing data into histograms, allows the analysis of uni and multivariate continuous, discrete ordinal and non-ordinal, as well as mixed distributions. In addition, methods have shown to be robust to detect changes on multivariate variable interactions.

As a consequence, the proposed methods have shown to be useful tools for data quality assessment focusing in the temporal variability dimension. This work has focused to the change monitoring and characterization on data distributions. The same concepts and methods can be applied to monitor other data quality features, such as monitoring the degree of missing, inconsistent, or incorrect data. These could be used to audit the quality of multi-centric or multi-user data gathering for research repositories, clinical trials, or claims data. Concretely, the latter are known to be far from perfect (Solberg et al, 2006), where these processes may be of special interest. Hence, the proposed methods can be used as exploratory data quality assessment solutions. Furthermore, as based on probabilistic metrics, they might also be used with quantitative decision making purposes. However further research is required to define these criteria.

5.6.2 Comparison with related work

Basic statistical methods, similarly to Shewhart control charts, have been used in the medical monitoring. E.g., laboratory systems have well established temporal quality controls based on the Levey-Jennings charts and Westgard rules (Westgard and Barry, 2010). Thus, a batch is considered Out-of-Control using basic statistics based on reference chemical reactives. On the other hand, other studies have used more complex change detection methods. Rodrigues et al (2011) proposed a method to improve the monitoring of cardiotocography signals using the memoryless fading window approach. Sebastião et al (2013) applied a Page-Hinkley change detection test combined with a time-weighted mechanism for the monitoring of depht anaesthesya signals. Similarly to the PDF-SPC, these studies focus to the monitoring of data itself, based on quality control references or in physiological signals.

On the other hand, Stiglic and Kokol (2011) proposed a method to facilitate the interpretation of changes in the performance of clinical diagnosis classification models by means of a bivariate analysis of class labels. Using the NHDS dataset, they found a change in the performance of models to predict chronic kidney disease by the end of year

2005. Their visual method provided the insights to confirm that the change was due to change in the ICD9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) title and description of concepts D403 and D404, related to the investigated disease. Additionally, a decrease in the performance measures was found by the start of year 2008. Interestingly, that change is correlated in time with the change in the codification of age found in this work.

In the generic change detection domain this is not the first study using non-parametric PDF distance measures for detecting changes. In their useful approach, Kifer et al (2004) proposed a relaxed total variation distance among PDFs for change detection. They discarded using information-theoretic distances claiming discrete distributions were needed. However, based on the Kullback-Leibler divergence they can be used on purely continuous and, as demonstrated in this work, even in mixed-types multi-variate distributions. Dasu et al (2009) and Sebastião et al (2010) did use the Kullback-Leibler divergence in their respective studies. However, the Kullback-Leibler neither satisfies the properties of a metric nor is bounded, as required in this study.

The higher dimensionality is a challenge for change detection methods. Several solutions have been proposed in the general change detection domain. Aggarwal (2003) proposed a method based in a physical model to measure the velocity of changes in probability masses of continuous data using KDE. To deal with higher dimensionality and facilitate the understanding on changes, he proposed picking sub-projections in which the greatest amount of change has occurred. Hrovat et al (2014) applied a strategy to detect relevant subgroups on which to make the analysis, with the purpose to detect temporal trends on biomedical data. Papadimitriou et al (2005) presented a method capable to find the key trends in a numerical multivariate time series. The method internally used principal component analysis (PCA), which may not result effective with multi-modal distributions nor reducing dimensionality including categorical data, both aspects generally present in biomedical data. The previously mentioned approach by Dasu et al (2009) measures changes between two PDFs embedded into a reduced structure using an extension of kd-trees. Thus, a distance metric needs to be defined between data points, what may be complicated when non-ordinal data is present. It can be deduced, hence, that dealing with non-ordinal discrete data may represent another challenge. This work deals with these problems using the synopsis on non-parametric discrete histograms.

Other approach for change detection suited to multi-modal distributions consists on monitoring cluster evolutions. Spiliopoulou et al (2006) presented the MONIC framework based on that idea, where different types of cluster changes are characterized. It is important to distinguish between such approaches and the temporal clustering presented in this work. Whilst the former clusters data within time windows, in this work what is clustered are the time windows.

5.6.3 Limitations

In high dimensions, the histogram synopsis method involves by default a larger probabilistic space. Hence, data points may become sparse, leading to ineffective distribution comparisons, such as larger PDF distances. Due to the heterogeneous conditions of biomedical data, the application of non-linear dimensionality reduction methods, such as ISOMAP (Tenenbaum et al, 2000), or solutions as exposed in related work may alleviate such problem. Evaluations on high-dimensional simulated changes may help studying these.

On the other hand, on continuous data, KDE provides smoother PDF estimations than raw histogram estimations. Although the fading window approach already smooths the obtained PDFs using past data, KDE improves the smoothing within the current window. However, sometimes using data models instead of raw data may hide other relevant information present at lower levels of probabilistic magnitude. An interesting example arises from the evaluated data. The histogram estimation of the age variable is shown in Figure 5.1(d) splitted by sex. It can be observed that in both sexes (separately in Figure C.1 and Figure C.2) there is a straight temporal gap beginning approximately at the age of 47, and continuing in a yearly basis. Considering that the NHDS represents the population of the United States, and that the first sample corresponds to year 2000, it is concluded that these population gap correspond to patients born around 1943, where the social effects of World War II reduced the birth rates. On the other hand, the use of methods to select the proper number of bins in histogram may be useful to overcome some of these issues, however, the proper number of bins may also vary through time, what may require further study to make compatible the on-line probability distance measurements as the support is changed.

Finally, we must recall that the IGT plot and PDF-SPC may not provide all the information related to the detection of heteroscedastic data, in contrast to the distribution heatmaps. Estimating distributions over time in a non-parametric approach, such as in this work, may facilitate detecting heteroscedastic behaviours. This is well observed in the distribution heatmaps, such as in Figure 5.1. However, the IGT plot and PDF-SPC represent the degree of difference among the temporal batches where, although the differences in heteroscedasticity will be captured, the source of heteroscedasticity is not shown (although it was not the purpose). Nevertheless, once the sources of heteroscedasticity are known, one could apply separate temporal variability analysis for each of them, or even apply a combined use of multi-source and temporal variability methods, such as proposed next in Section 6.1.2.

5.6.4 Future work

It has been observed in the examples that a recurrent change is related to a 12-month periodic seasonal effect. During the experiments developed in this work, it was observed that applying a simple non-weighted sliding window scheme with a window size of 12 months completely removed such effect on the resultant monthly PDF estimations (see Figure C.3). That effect may result useful for the detection of gradual changes, since higher frequent, recurrent changes which may hinder the former are removed. However, the long-term smoothing may cause abrupt changes not to be accurately detected. Hence, two interesting future work topics arise. First, automatically detecting the period of recurrent changes, e.g., based on signal processing methods. Second, using ensemble change detection models combining different window schemes focused to specific types of changes.

The study of proper dimensionality reduction methods for effectively measuring PDF distances on high dimensions is also an important future work. This can be complemented with methods to select appropriate variable or sample subgroups on which to make the analysis. In addition, the maintenance and compatibility through time of these methods to reduce the problem complexity can be studied.

Regarding to the temporal characterization and subgroup discovery methods, it is open as further work their improvement based on incremental approaches. Hence, incremental clustering (Rodrigues et al, 2008) and MDS (Brandes and Pich, 2007) methods could be used with such a purpose. On the other hand, the use of complementary methods to optimise the efficiency of the projections, such as Self-Organizing maps (Kohonen, 1982), may be studied.

Another interesting future work is to apply functional data analysis methods (Ramsay and Silverman, 2005) to model the probabilistic temporal evolution of data on the information-geometric statistical manifold. They may provide smoothed tendency curves on which to characterize and measure changes.

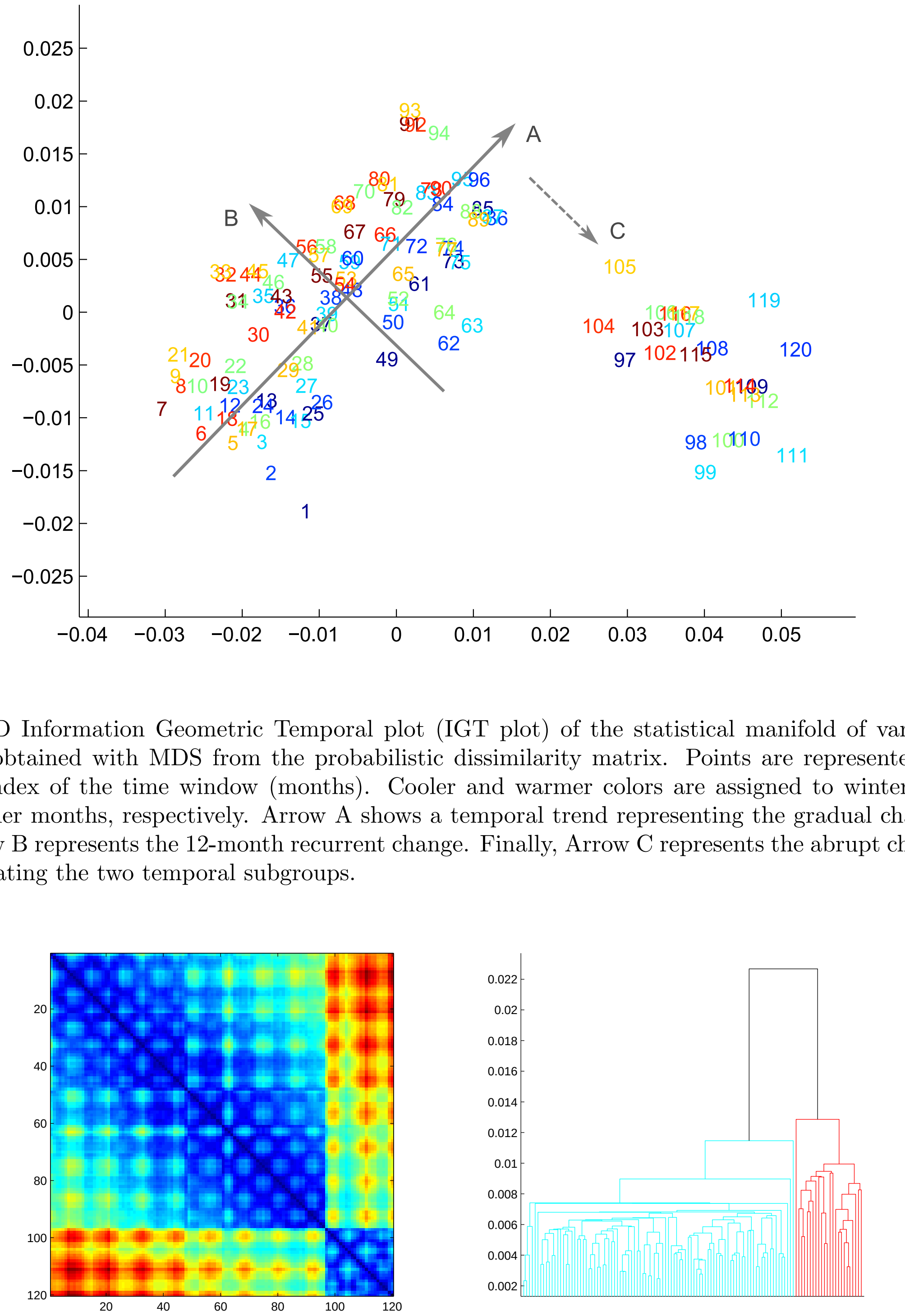
Finally, the combination of the temporal variability methods presented in this work with metrics for the probabilistic *multi-source variability* among multiple sources of biomedical data (Sáez et al, 2014b), will lead to a future study aiming to a probabilistic spatio-temporal data quality assessment.

5.7 Conclusion

The probabilistic methods presented in this work have demonstrated their feasibility for the change detection, characterization and subgroup discovery of temporal biomedical data. The changes present in the evaluated and simulated NHDS datasets have been successfully detected, in addition, with a probabilistic interpretation, as provided by the proposed PDF-SPC and information-geometric projection methods. Further studies will be made to confirm the generalisation of the methods.

As part of a data quality assessment, the proposed methods can facilitate the data understanding and lead to better decisions when developing knowledge discovery studies, either on-line or off-line, based on these data. Used as an exploratory framework, they permit visualizing the temporal variability of large healthcare databases in an interpretable and rapidly manner. In addition, methods are built to be comparable among different domains, hence, they may be used as part of a biomedical data quality auditory process. This is an important subject, as poor levels of data quality may have direct consequences on patient care (Aspden et al, 2004) as well as in the biomedical research processes (Weiskopf and Weng, 2013; Sáez et al, 2014b). This work has demonstrated that data stream and change detection methods can be successfully applied in the biomedical data context, thus, further studies can still be made to analyse the impact that a temporal variability assessment can provide to real, in-production healthcare repositories.

Finally, this work has contributed to the generic change detection field of study in two aspects. First, the extension of the widely accepted SPC method to the monitoring of changes in non-parametric PDFs based on information-theoretic distances. And second, the novel change characterization method based on information geometry. It is important to emphasize the contribution to the state-of-the-art of this method. In this work, it has demonstrated possibilities which have not received proper attention in the literature yet, such as discovering temporal subgroups or characterizing the direction and length of changes through the series of time-windows in the statistical manifold. However, a lot of new possibilities are opened, standing as the first step of a promising line of research in change detection.

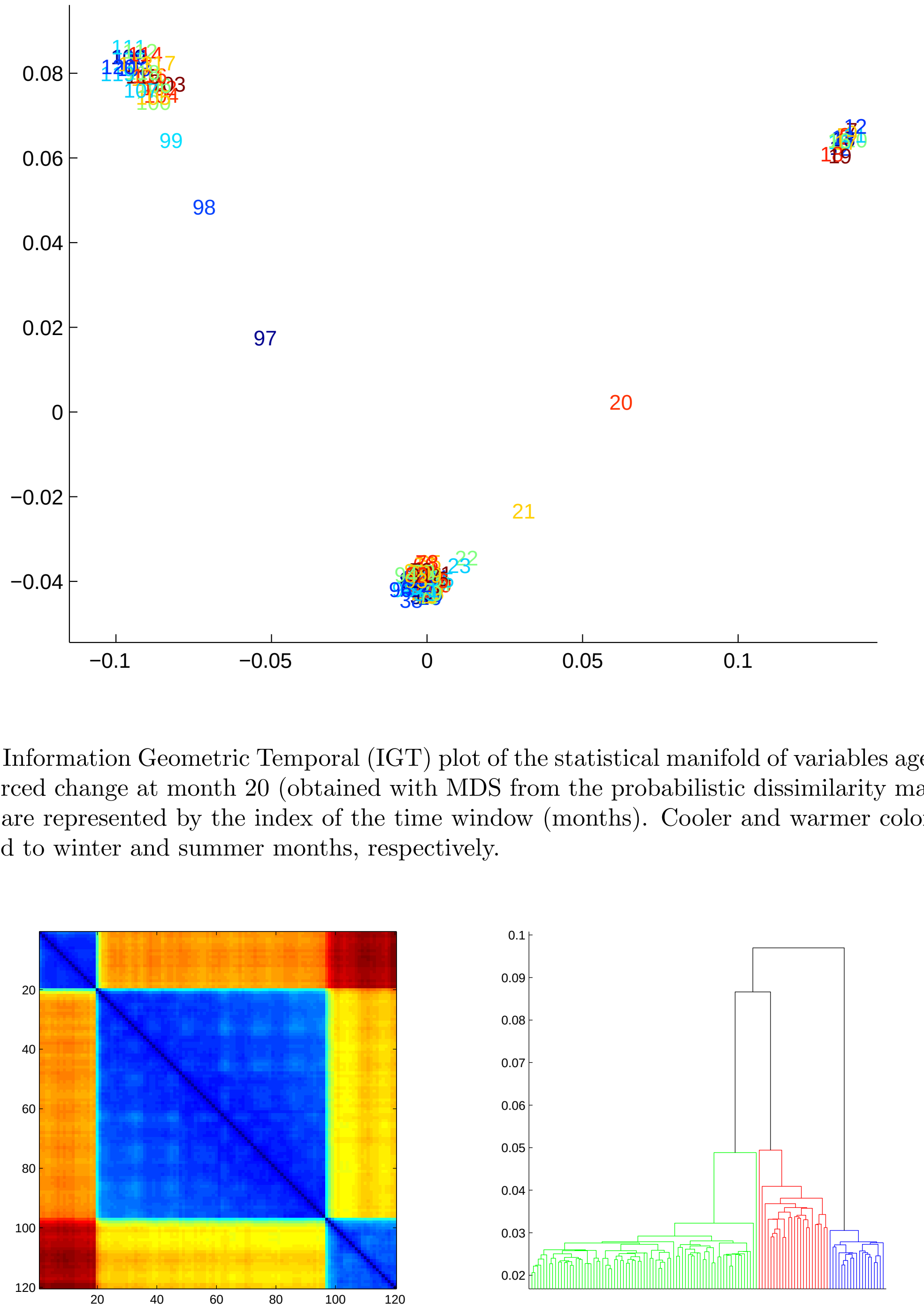


(a) 2D Information Geometric Temporal plot (IGT plot) of the statistical manifold of variable age, obtained with MDS from the probabilistic dissimilarity matrix. Points are represented by the index of the time window (months). Cooler and warmer colors are assigned to winter and summer months, respectively. Arrow A shows a temporal trend representing the gradual change. Arrow B represents the 12-month recurrent change. Finally, Arrow C represents the abrupt change separating the two temporal subgroups.

(b) Dissimilarity matrix heat map

(c) Dendrogram for temporal subgroups

Figure 5.3: Change characterization and temporal subgroup discovery on the age variable. Change characterizations are shown by temporal trends on the projection. Two subgroups are clearly observed in the approximated statistical manifold (a) and dissimilarity matrix (b), which were confirmed with a complete linkage hierarchical clustering (c).



(a) 2D Information Geometric Temporal (IGT) plot of the statistical manifold of variables age+sex with forced change at month 20 (obtained with MDS from the probabilistic dissimilarity matrix). Points are represented by the index of the time window (months). Cooler and warmer colors are assigned to winter and summer months, respectively.

(b) Dissimilarity matrix heat map

(c) Dendrogram for temporal subgroups

Figure 5.4: Change characterization and temporal subgroup discovery on the joint age and sex variables with forced change at month 20. Change characterizations are shown by temporal trends on the projection. Three subgroups are clearly observed in the approximated statistical manifold (a) and dissimilarity matrix (b), which were confirmed with a complete linkage hierarchical clustering (c).

Chapter 6

Applications to case studies

This chapter describes the application of the multi-source and temporal variability methods developed in this thesis to several case studies. The main case is the Public Health Mortality Registry of the Region of Valencia, Spain. This is a large multi-source dataset, which served as the validation benchmark on which both multi-source and temporal variability methods were systematically applied. Other case studies in this thesis include the Public Health Cancer Registry of the Region of Valencia, a National Breast Cancer multi-source dataset and an In-Vitro Fertilization dataset. Several variability findings obtained by the methods and their causes are described in this chapter. These results validate the use and usefulness of the methods as part of DQ assessment procedures in real scenarios.

The introductory notes of this chapter and the Mortality Registry case study are accepted in scientific publication by Sáez et al (2016)—thesis contribution P5.

6.1 Introductory notes

6.1.1 Summary of the applied methods

The methods used in the present chapter fall into two groups, namely those for assessing multi-source variability (which development is described in Chapter 4) and those for assessing temporal variability (which development is described in Chapter 5). Both methods are based on the comparison of probability distributions of the variables among different sources (e.g., sites) or over different time periods (e.g., months). In concrete terms, the comparisons are made by calculating the information-theoretic probabilistic distances (Section 2.2.2) between pairs of distributions. These comparisons offer a robust alternative to classical statistical tests, where there may not be appropriate (see Chapter 3).

The assumption made in the application of the methods is that in a repository with low variability, differences among distributions would be small whereas different or anomalous data distributions would mean higher variability. To facilitate the comparability among studies, the maximum distance is limited to one in all the methods. This indicates that when a distance between two distributions is one, the two distributions do not share common values. It is worth mentioning that probabilistic distances can be applied either to measure differences either in numerical data, such as ages, or categorical data, such as coded values, in a multivariate setting and are independent of sample size.

Data distributions in a repository are a representation of reality, which is measured or observed and then registered in an information system. Any change in the original real-world information or in data acquisition and processing—including changes due to other systematic DQ problems, such as missing data—will have a much greater chance of being reflected in the data probability distributions: the metrics and visualizations proposed in this thesis, being probability based, are thus likely to capture most of these changes.

Appendix D provides basic, illustrative examples of the methods to assess both multi-source and temporal variability and complement the descriptions given below.

Multi-source variability

The multi-source variability method (Chapter 4) involves constructing a geometric figure (referred to as a multi-source variability simplex), the points of which represent data sources and the lines that join the points (the lengths of the lines) represent the measured distances between the distributions of the data sources. For example, a triangle represents three data sources. Generalizing to any number of data sources, the adequate geometric figure is a simplex, the generalization of a triangle to multiple dimensions. The centroid of the simplex represents the hidden average of the distributions of sources in the repository. Based on this simplex, we get to the next two metrics and exploratory visualization.

Source Probabilistic Outlyingness (SPO) metric: This metric measures the dissimilarity of the distribution of a single data source to the global average, which serves to highlight anomalous data behavior. From the multi-source variability simplex, we calculate the SPO of each source based on the distance between the point that represents a given source and the simplex centroid.

Global Probabilistic Deviation (GPD) metric: This metric shows the degree of global variability among the distributions of sources in a repository, as an estimator equivalent to the standard deviation among distributions. It is calculated based on the mean of the distances between each point that represents a source and the simplex centroid. Both the GPD and SPO metrics are bounded by zero and one. Further, using the simplex centroid as a reference hidden average distribution avoids the need for a reference gold-standard data set.

Multi-Source Variability (MSV) plot: This is a visualization of multi-source variability representing the simplex figure as a two-dimensional (2D) plot with its two axes representing its two most relevant dimensions, which we named D1-simplex and D2-simplex. In the visualization, data sources are shown as circles in which the distance between two circles represents the distance between their probability distributions. As a consequence, data sources with similar distributions are grouped together, whereas those with different or anomalous distributions are positioned far away. As an alternative visualization to that shown in Chapter 4, the radius of a given circle is made proportional to the number of cases in the data source and the color of each circle indicates the SPO of the source.

Temporal variability

The temporal variability method (Chapter 5) involves comparing the distributions through different batches of data in the repository, each batch representing a user-specified interval (weeks, months, years, etc.).

Information Geometric Temporal (IGT) plot: This provides an exploratory visualization of the temporal evolution of data. The idea behind it is similar to that for the MSV plot. The temporal batches are laid out as a 2D plot while conserving the dissimilarities among their distributions: based on the positions of the temporal batches, we can appreciate the evolution of the distributions in the repository over time. Thus, the IGT plot helps in uncovering temporal trends in the data (as a continuous flow of points), abrupt changes (as an abrupt break in the flow of points), recurrent changes (as a recursive flow through specific areas), conceptually related time periods (as grouped points), and punctual anomalies (as isolated outlying points). To facilitate the interpretation, temporal batches are labeled to show their date, given suitable colors (warm colors for summer and cool colors for winter, for example), and supported by a smoothed timeline path.

Probability Distribution Function Statistical Process Control (PDF-SPC) algorithm: This provides an automated statistical process control (SPC) for monitoring changes in data distributions, similar to classical SPC methods. In classical SPC methods, a numerical parameter under control is monitored and kept within some limits; for example, the results of reactivity controls in laboratory systems are periodically tested to check whether they are within the acceptable limits of error (Westgard and Barry, 2010). Similarly, the purpose of PDF-SPC is to monitor the variability of data distributions through consecutive temporal batches. This is done by monitoring an upper confidence interval (e.g., one standard deviation) of the accumulated distances of temporal batches to a reference distribution (initially the first batch). According to the magnitude of the current confidence interval, the degree of change of the repository is classified into three states: in-control (distributions are stable), warning (distributions are changing), and out-of-control (recent distributions are significantly dissimilar to the reference, leading to an unstable state). Warning states can be false alarms if the distances get closer to the reference once again, thus going back to the in-control state. However, when an out-of-control state is reached, a significant change is confirmed and the reference distribution is set to the current. The results of the PDF-SPC algorithm are made visible in a control chart, which plots the current distance to the reference, the mean of the accumulated distances, and the upper confidence interval being monitored and indicates the warning and out-of-control states as broken or continuous vertical lines, respectively.

Temporal heat maps: As a support to the temporal variability methods, we proposed the use of temporal heat maps for absolute (counts) or relative (probability distributions) frequencies. These are 2D plots where the X axis represents the time, the Y axis represents a possible data value or range of values, and the color of the pixel at a given (X,Y) position indicates the frequency at which value Y was observed on date X. These heat maps facilitate a rapid and broad visualization of how the values of a variable evolve over time.

6.1.2 Additional method combining multi-source and temporal variability

Additionally, for the case studies we developed a new method based on monitoring multi-source variability over time. Similarly to the temporal variability methods, this new method involves calculating the SPO and GPD metrics, or the MSV plot, through continuous temporal batches and plotting their results.

SPO and GPD monitoring: The GPD and SPO are scalar metrics obtained from a set of distributions, independently of the data acquisition times. As a consequence, given a series of temporal batches, we can obtain for each the GPD and SPO metrics. The resultant metrics will be comparable over time, and they can be translates to a time plot for their monitoring.

MSV plot monitoring: With the purpose of monitoring the MSV plot through temporal batches the first choice would be obtaining their corresponding simplices at each iteration. However, we found that with that solution the resultant visualizations were not smoothly transited, difficulting their comparability over time. The reasons behind that problem are two. First, the MDS algorithm applied to calculate the simplex provides the optimum projection only for the points received as input. And second, the resultant projection is independent to rotation, then, infinite equivalent variations of results are possible.

As a consequence, we opted for a solution which provided smoothed results. This consisted in two steps. First, we used at each iteration as the input for MDS both the current multi-source distributions and all the previous ones. In this manner, the resultant simplex will obtain a simplex projecting all the sources for all temporal batches. Next, the second step consists in dividing the projected points according to their temporal batch, what leads to one simplex projection per batch. In this approach, the MDS algorithm optimizes the dissimilarities among the distributions of all the sources through all their temporal batches, hence, the coordinates of the point corresponding to a specific source will be comparable to their inner dissimilarity through time, leading to smoothed overall results for all sources.

6.2 Mortality Registry of the Region of Valencia

6.2.1 Materials

The above methods were applied to the Public Health Mortality Registry of the Region of Valencia (MRRV), an autonomous region of Spain along the Mediterranean coast. The repository comprises the records related to a total of 512 143 deaths that occurred between 2000 and 2012 (inclusive), disaggregated by 24 health departments (6.1) covering 542 cities and towns with a total of 4.7 million inhabitants on average, representing 11% of the population of Spain. The repository includes the variables that make up the Spanish National Medical Death Certificate, an official paper document completed by a physician after the death of a person, according to the recommendations of the World Health Organization (WHO) (WHO, 2012). Any information that may disclose the identity of the person was removed before the analysis.

The studied variables include demographic information, the groups of sequential causes leading to death (multiple causes), and the basic cause of death (Table 6.2). Multiple causes include initial, intermediate, immediate and contributive causes. For each group, up to three values are entered depending on the number of causes. Further on, empty values up to the three possibilities will be labeled as ‘not applicable (NA)´. The basic cause of death is the official cause, which is the one taken into account for national and international mortality statistics and generally coded afterwards by specialist staff based on the multiple causes listed in the certificate.

According to the WHO recommendations, for facilitating statistical analysis and comparison of this work to other international studies, the causes of death were recoded using the WHO International Classification of Diseases (ICD) version 10 Mortality Condensed List (WHO, 2009), which condenses the full range of ICD three-character categories into 103 manageable items. Because this list brings together both the top-level ICD chapters and their subgroups of diseases, the chapter-level classifications were discarded to avoid duplication and to facilitate proper statistical distribution. Accordingly, a total of 92 unique causes of death (plus an additional category, namely NA) were used in the present study (Section E.1, Appendix E). Deaths that occurred outside the Region of Valencia during this period (totaling 6,816) were excluded, leaving us finally with 505,327 entries. The Consolidated Standard of Reporting Trials (CONSORT) diagram of the study is shown in Figure 6.1. Additionally, a map showing the 24 health departments and tables of sample sizes are included in Appendix E.

6.2.2 Results

The results of applying the methods to the MRRV repository are shown below following a discovery process that led to four types of findings.

Temporal anomalies

We first analyzed the temporal variability of the multivariate MRRV repository as a whole using IGT plots. To simplify the analysis, all the variables, including both

Table 6.1: List of Health Departments of the Region of Valencia (version 2010, as used in this study)

Short name	Acronym	Full name of Health Department	Province
Vinaròs	Vi	Departamento de Salud de Vinaròs	Castellón
Castellò	C	Departamento de Salud de Castellón	Castellón
LaPlana	LP	Departamento de Salud de la Plana	Castellón
Sagunt	S	Departamento de Salud de Sagunto	Castellón - Valencia
ClínicMir	CM	Departamento de Salud de Valencia - Clínico - Malvarrosa	Valencia
ArnauLlíria	AL	Departamento de Salud de Valencia - Arnau de Vilanova - Llíria	Valencia
Manises	M	Departamento de Salud L'Horta Manises	Valencia
Requena	R	Departamento de Salud de Requena	Valencia
VGral	VG	Departamento de Salud de Valencia - Hospital General	Valencia
Peset	P	Departamento de Salud de Valencia - Doctor Peset	Valencia
LaRibera	LR	Departamento de Salud de la Ribera	Valencia
Gandía	G	Departamento de Salud de Gandía	Valencia
Dénia	D	Departamento de Salud de Dénia	Alicante
XàtivaOnt	XO	Departamento de Salud de Xàtiva - Ontinyent	Valencia
Alcoi	Ac	Departamento de Salud de Alcoy	Alicante
MarinaB	MB	Departamento de Salud de la Marina Baixa	Alicante
SantJoan	SJ	Departamento de Salud de Alicante - San Joan d'Alacant	Alicante
Elda	El	Departamento de Salud de Elda	Alicante
Elx	E	Departamento de Salud de Elche - Hospital General - Crevillent	Alicante
AGral	AG	Departamento de Salud de Alicante - Hospital General	Alicante
Orihuela	O	Departamento de Salud de Orihuela	Alicante
Torrevieja	T	Departamento de Salud de Torrevieja	Alicante
València	V	Valencia ciudad	Valencia
Alacant	A	Alicante ciudad	Alicante

numerical and categorical data, were combined using the principal component analysis (PCA) dimensionality reduction method. Figure 6.2 (a) shows the IGT plot for 2000–2012 giving the distributions of monthly temporal batches. The distributions from January to March 2000 (arrows a, b, and c) are located at anomalous positions with respect to the distributions for other months and according to the time flow. This indicates anomalous behavior of the data for these three months. Drilling down to specific variables, the anomaly was found in all multiple causes as well. The associated heat map of the temporal distribution of these variables helped in uncovering the punctual increment on unfilled data for these months, reaching almost 100% in some variables (e.g., see Section E.3, Appendix E).

To avoid a possible bias in the results pertaining to the year 2000, the first decision in the procedure of assessing DQ was to exclude the entire year, given the difficulty in recovering all the missing data.

Table 6.2: Studied variables of the Public Health Mortality Registry of the Region of Valencia

Variable	Description	Type
Age	Age in years at the time of death	Numerical integer
Sex	Sex of the person	Categorical {Male, Female}
Basic cause	Basic cause of death	ICD-10 List 1 code
ImmediateCause[1,2,3]	Disease or condition directly leading to death (one to three options)	ICD-10 List 1 code
IntermediateCause[1,2,3]	Morbid conditions, if any, giving rise to the above cause (one to three options)	ICD-10 List 1 code
InitialCause[1,2,3]	Disease or lesion that initiated the process that eventually resulted in the death (one to three options)	ICD-10 List 1 code
ContributiveCause[1,2,3]	Other significant conditions contributing to the death but not related to the disease or condition that caused death (one to three options)	ICD-10 List 1 code
Health Department	Health department the person was assigned to (associated with the city of residence)	Discrete code

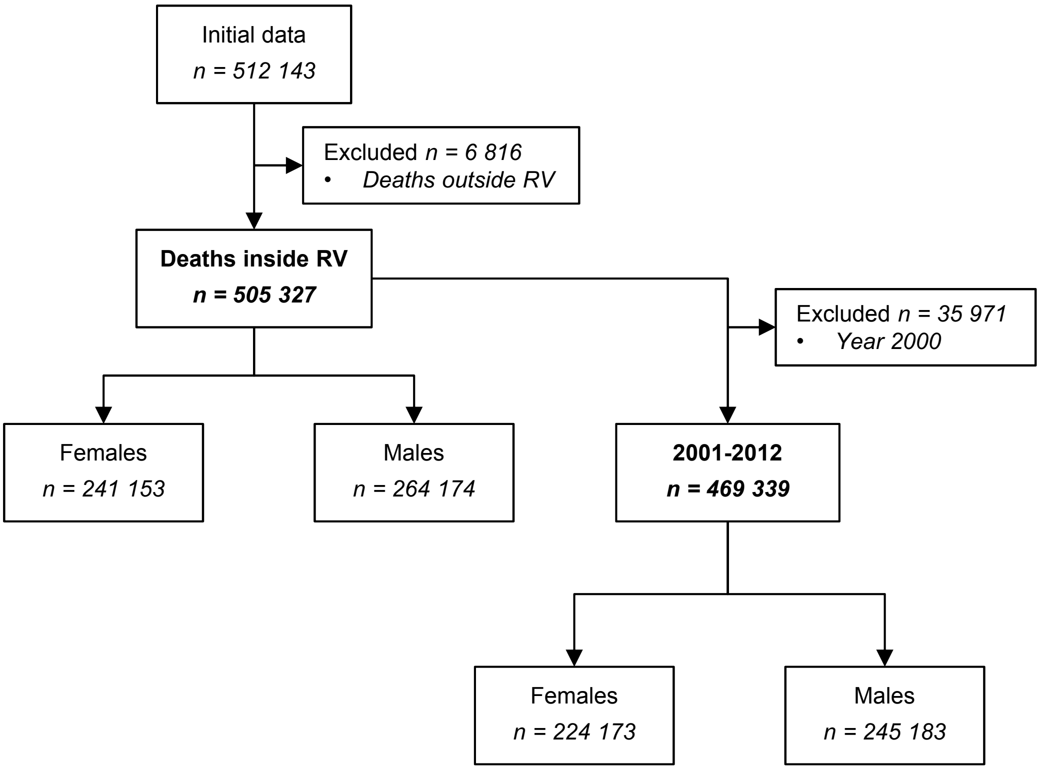


Figure 6.1: CONSORT flow diagram of the case study of the Mortality Registry of the Region of Valencia (RV)

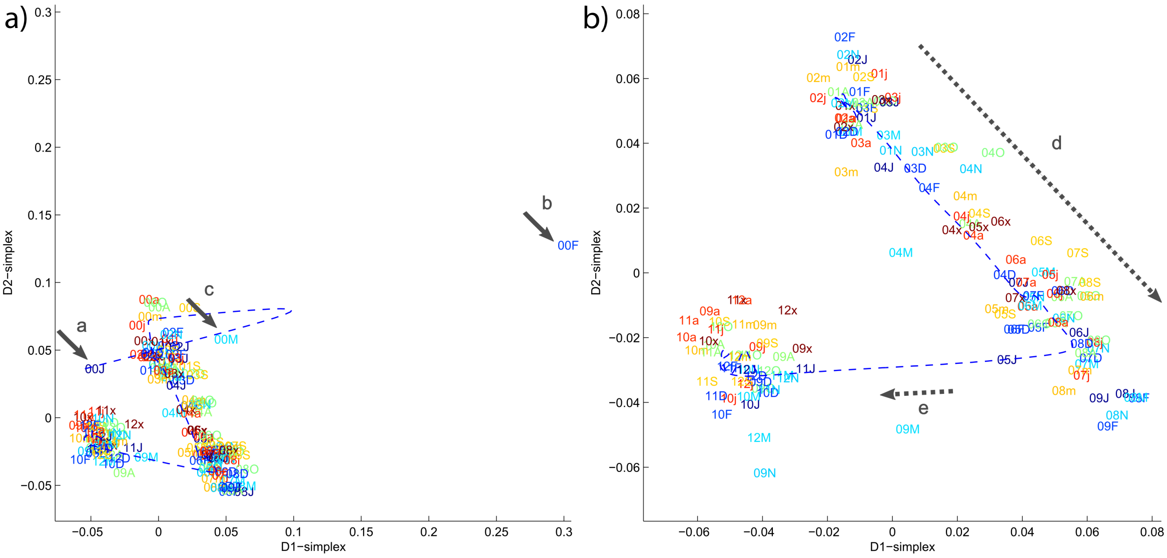


Figure 6.2: IGT plots of the multivariate repository on monthly basis. Each point represents one batch of the repository labeled with its date in 'YYM' format (YY: the last two digits of the year, M: the month as given in the list of abbreviations at the end), and the distances among them represent the dissimilarity in their distributions. a) The period 2000–2012, where the months January to March 2000 (arrows a, b, and c) are at anomalous positions according to the time flow. b) The period 2001–2012, after discarding the data for 2000. A gradual conceptual change is seen from the start until 2009 (arrow d), at which point the change is abrupt (arrow e), splitting the repository into two temporal subgroups. The cool (blues) and warm (yellows and reds) colors indicate winter and summer months, respectively.

Temporal subgroups

Figure 6.2 (b) shows the IGT plot of the multivariate MRRV repository in 2001–2012. The flow of points is continuous through the timeline (arrow d) until February 2009, indicating a gradual change in their distributions. An abrupt change in March 2009 (arrow e) then splits the repository into two temporal subgroups, i.e., conceptually-related time periods. Additionally, a yearly seasonal component can be observed, especially in the latter subgroup, based on the color temperature of the months.

Figure 6.3 shows the PDF-SPC chart for 2001–2012. After a transient state (2001), the change is gradual, corresponding to a gradual increase in the distribution distance to the latest reference month, alerting two warning states around 2004 (broken vertical lines) until the accumulated threshold is reached in 2008 leading to an out-of-control state (solid vertical lines). The abrupt change in 2009 was detected by the method and confirmed afterward.

Drilling down to specific variables, we found that the abrupt change in 2009 was also present for most groups of the causes of death. For example, Figure 6.4 (a) shows the IGT plot of *IntermediateCause1*, where the change is observed in March 2009, plus an additional abrupt change in 2011, a gradual change, and a seasonal effect. The temporal heat maps of the variables (Section E.3, Appendix E) uncovered a major change in the number of specified causes in 2009. This situation is summarized in Figure 6.5.

To check whether such an abrupt change was solely due to the number of specified

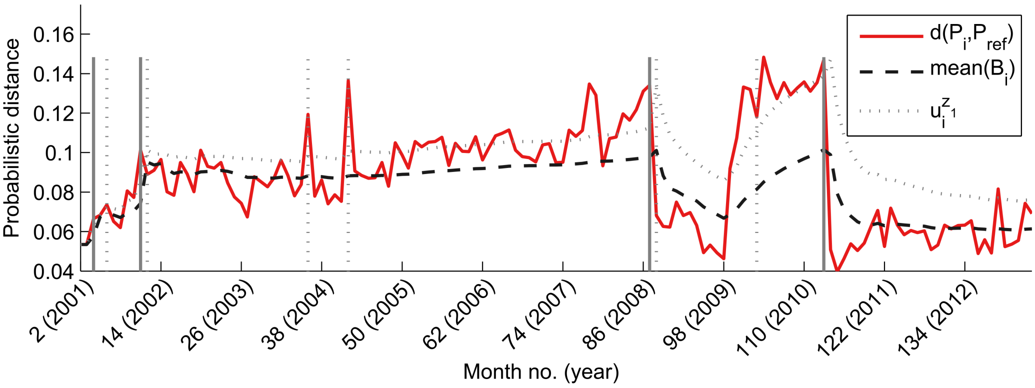


Figure 6.3: PDF-SPC monitoring of the variability of the distribution of the entire repository on a monthly basis. The chart plots the current distance to the reference $d(P_i, P_{ref})$, the mean accumulated distance (mean(B_i)), and the upper confidence interval being monitored $u_i^{z_1}$ and indicates the warning and out-of-control states as broken or continuous vertical lines, respectively. After a transient state (2001), a gradual change is seen, alerting two warning states around 2004, until the threshold is reached in 2008, leading to an out-of-control state, which re-establishes the reference distribution. The abrupt change in 2009 is captured by the metric and confirmed afterward.

causes, we ignored the NA category in the distributions and focused on the 92 unique codes in ICD-10 List 1. However, the change persisted for most of the variables even after recomputing (Figure 6.4, (b)), indicating that the frequencies of causes of death changed abruptly as well (although to a small extent). Figure 6.4 (c) shows the temporal heat map of the distribution of *IntermediateCause1* without the NA category. Hence, in 2009, the frequencies of 'hypertensive diseases´, 'chronic lower respiratory diseases´, and 'diabetes mellitus´ increased whereas those of 'symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified´ decreased, among others. However, in 2011, some of those frequencies were re-adjusted.

As a consequence, the separation of the repository into two temporal subgroups, up to 2009 and from 2009 onward, gives the first hint that statistical analyses or models that treat the entire span as one may not be concordant, given the abrupt differences in their data distributions. Consequently, in some of the further steps, the two subgroups were analyzed separately.

Departmental anomalies

We next assessed the variability among different health departments, starting with anomalies, if any, in data from individual departments. Figure 6.6 (a) shows the SPO monitoring of the multivariate MRRV repository on yearly basis for the period 2001–2012. The health department of Requena, in the western part of the region, showed a large SPO, indicating an outlying distribution. Besides, the health department of Torrevieja, in the southern part of the region, also increased its SPOs during 2005–2009.

Further scrutiny led to the splitting of the set of variables into two subgroups: one classifying individual deaths by *Age*, *Sex*, *BasicCause* and other representing deaths as registered in the Certificate due to multiple causes. The latter subgroup behaved

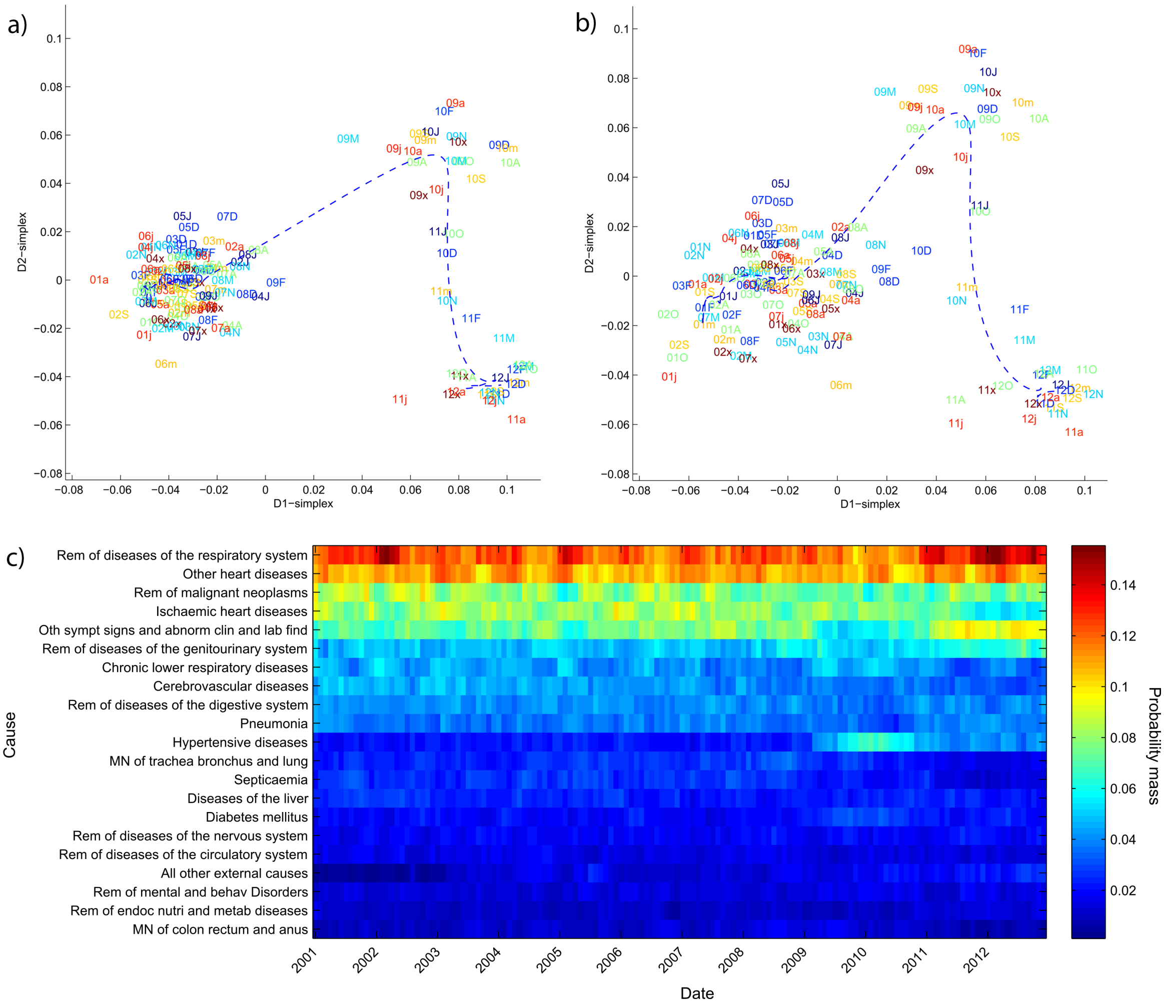


Figure 6.4: IGT plots (a, b) and temporal heat map of distribution (c) of *IntermediateCause1* for men in 2001–2012 on monthly basis. Each point represents one batch of the records labeled with its date in 'YYM' format (YY: last two digits of the year, M: the month as given in the list of abbreviations at the end), and the distances among them represent the dissimilarity in their distributions. The IGT plots were calculated considering (a) and discarding (b) unfilled values (NAs). The heat map shows the evolution of the probability distribution for 21 most prevalent causes after discarding the NA category. The three main temporal subgroups seen in both the IGT plots (split by months, namely 09M and 11J) are associated with the changes in the patterns of the frequencies of causes shown in the heat map for 2009 and 2011.

the same way as the entire group, with a predominant SPO in Requena, followed by Torrevieja and Orihuela. In contrast, in the former subgroup we found a predominant SPO in the departments of Torrevieja and Valencia. Figure 6.6 (b) and (c) show the MSV plots of the two subgroups in 2008, showing interdepartmental dissimilarities.

Drilling down to individual variables, we found Requena as outlier for *Contributive-Causes[1,2,3]*. However, the anomaly disappeared after discarding the category NA. We then analyzed the number of filled causes by the departments and found that Requena was the department that had filled the maximum number of contributive causes (Section E.5, Appendix E). We also found that Torrevieja was outlier for the age at death, being the opposite of Requena (Section E.5, Appendix E).

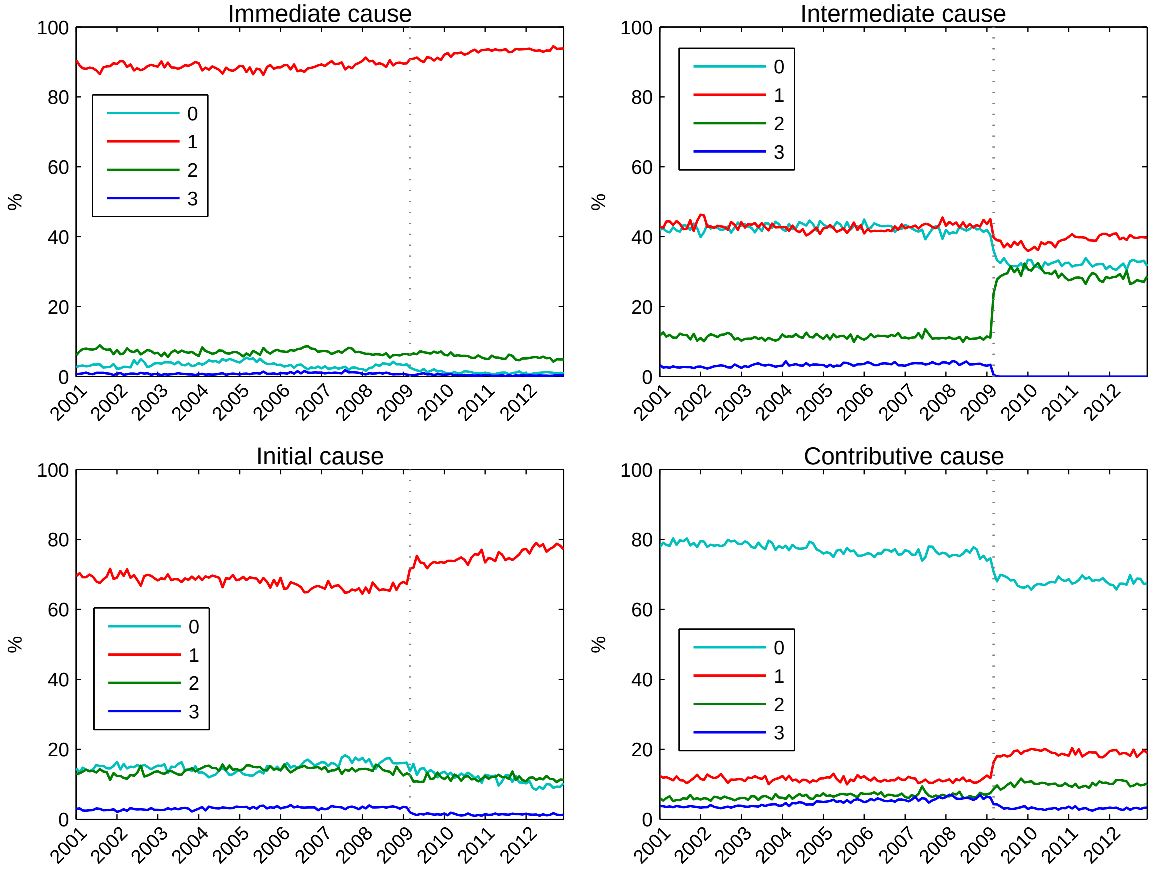


Figure 6.5: Number of filled causes by group of causes in the period 2001-2012 for both sexes. For each case the Mortality Registry can contain from zero to three filled items by group of causes, according to what was specified by the physician on each death. The figure shows the percentage of times each number of causes was filled for each group in a monthly basis. Besides the evident differences among groups, the main finding is the abrupt change in the number of specified causes in 2009, especially in Intermediate, Initial and Contributive causes

Departmental subgroups

The existence of source subgroups, i.e., groups of sources with similar probability distributions, was addressed next. The MSV plots uncovered a multi-site subgroup formed by most departments in the province of Alicante, mainly found in *ImmediateCause1*, *IntermediateCause1* (Figure 6.7), and *InitialCause1*. Discarding the category NA, the subgroup was not present in *InitialCause1*; however, it still was present in *ImmediateCause1* and *IntermediateCause1*. The subgroups were empirically confirmed using clustering algorithms based on the dissimilarity matrix of interdepartmental distribution distances obtained from the method (Section E.6, Appendix E).

The change to the death certification in 2009 can be seen in Figure 6.7 as a global change affecting all data points in the last batch (2009-2012), but not necessarily equally.

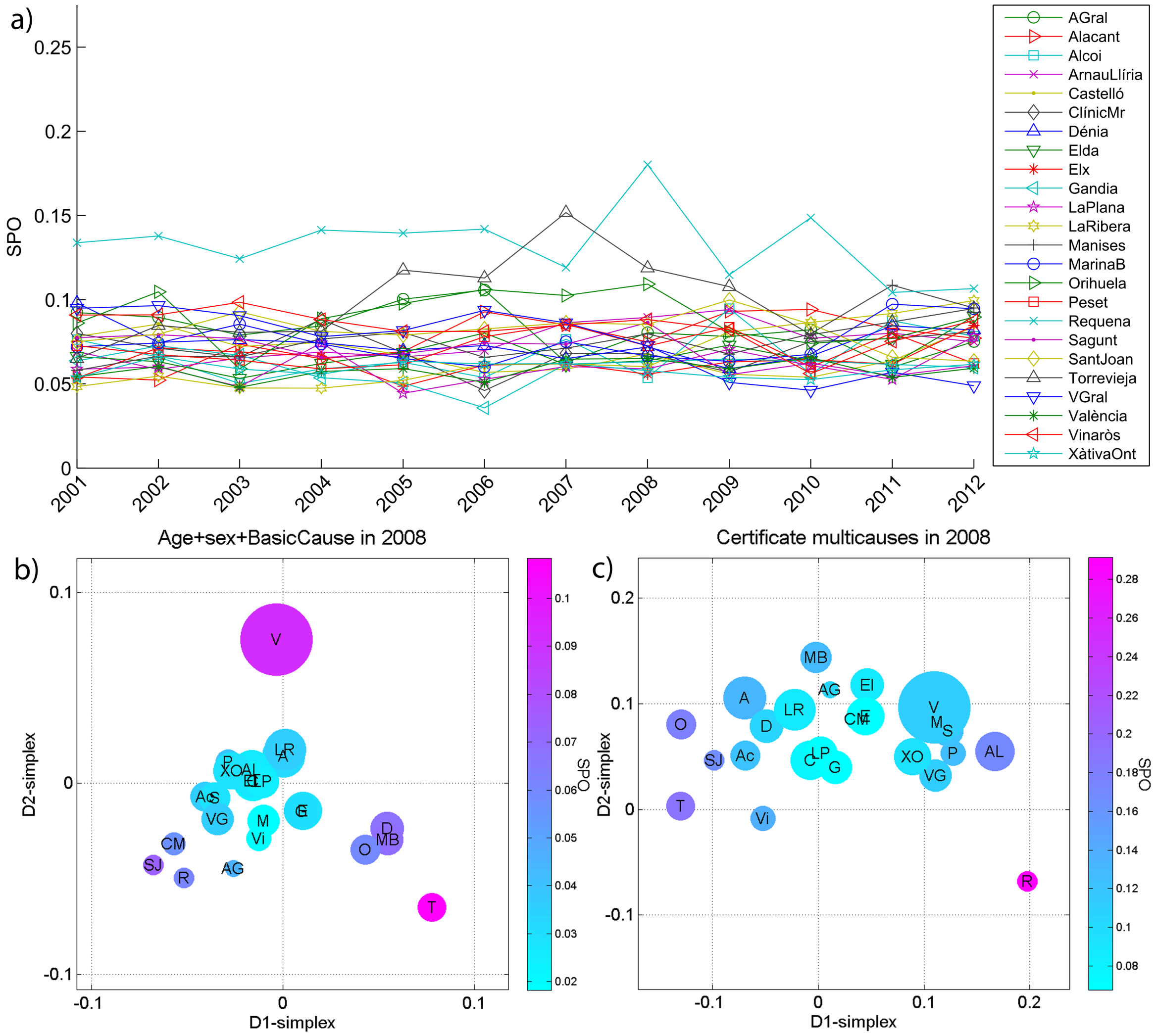


Figure 6.6: Monitoring of the departmental anomalies based on the distribution of all variables in the repository during 2001-2012. We used SPO monitoring (a), variability among the distributions of the health departments in 2008 based on multivariate combinations *Age*, *Sex*, *BasicCause* (b), and multiple causes (c) with MSV plots. Circles represent the health departments (see the key to the names at the end), their color represents the source SPO, and their size reflects the sample size.

Global concordance among Departments

Finally, we measured the GPD for the main groups of causes-of-death among all the Health Departments in the Mortality Registry, dividing the period of study in three batches of four years. The GPD was measured separately for females and males, and considering and not considering the unfilled (NA) values. The latter was done with the purpose to focus on the causes of death their selves, avoiding the effect that the different number of unfilled causes among Departments may introduce.

The results are shown in Figure 6.8. The series show the evolution through time of the GPD metric in the different groups of causes-of-death. We first note that when not considering the unfilled (NA) values (right column) the variability among Departments is reduced in all groups. This may indicate that either the causes-of-

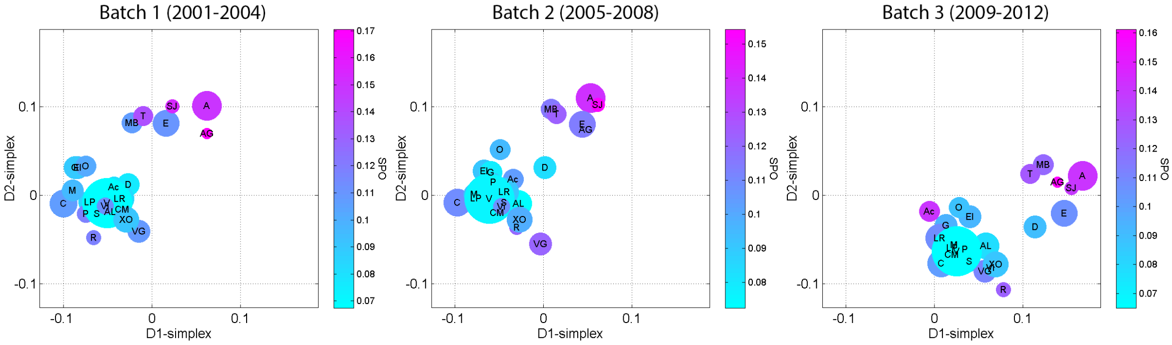


Figure 6.7: Variability of *IntermediateCause1* among the distributions of the health departments over time in batches of four years using MSV plot monitoring. Circles represent the health departments (see the key to the names at the end), their color represents the source SPO, and their size reflects the sample size. A subgroup formed by most departments in the province of Alicante is at upper right part throughout. Besides, the change to the death certification in 2009 can be seen as a global change affecting all data points in the last batch (2009-2012), but not necessarily equally.

death or the healthcare or death certifying practices are becoming more similar among Departments. However, when considering unfilled data (left column), the variability remains stable, or is increased or reduced, depending on the group. This indicates that there exist differences among health Departments in the number of specified causes, related to different certification practices respect to the unfilled causes.

6.2.3 Discussion

Table 6.3 summarizes the main findings and their causes—the result of applying multi-source and temporal variability assessment methods to the MRRV repository. Such a table may constitute a form of feedback for the management of DQ in repositories of biomedical data. First, the table may serve as a reference to avoid any problem or bias caused by multi-source or temporal variability in the data to be reused. Second, the table serves as feedback for improving the processes of data acquisition and repository maintenance and for preventing future problems related to DQ.

Probably the most important finding from this exercise is the abrupt change in 2009, leading to abrupt variations in the number of specified causes and in the incidence of some causes of death described above. This change coincides with the redesign of the National Certificate of Death in 2009. The new certificate was intended to meet the WHO recommendations to a greater extent while retaining the earlier structure of the certificate as much as possible. Two modifications to the certificate probably account for the abrupt change in 2009, namely (1) the use of a row of boxes, each to be filled with one letter, instead of blank lines that allowed continuous writing, and (2) renaming the field ‘Intermediate cause´ as ‘Antecedent cause´ and providing one more line for the entry. The first modification may have reduced the chances of filling more than one cause and encouraged filling at least one. The second modification probably increased the frequency of cases in which two intermediate causes were entered but, at the same time, limited the entries to only two causes—the option of entering a third cause was never used (Figure 6.5). Additionally, the renaming caused some physicians

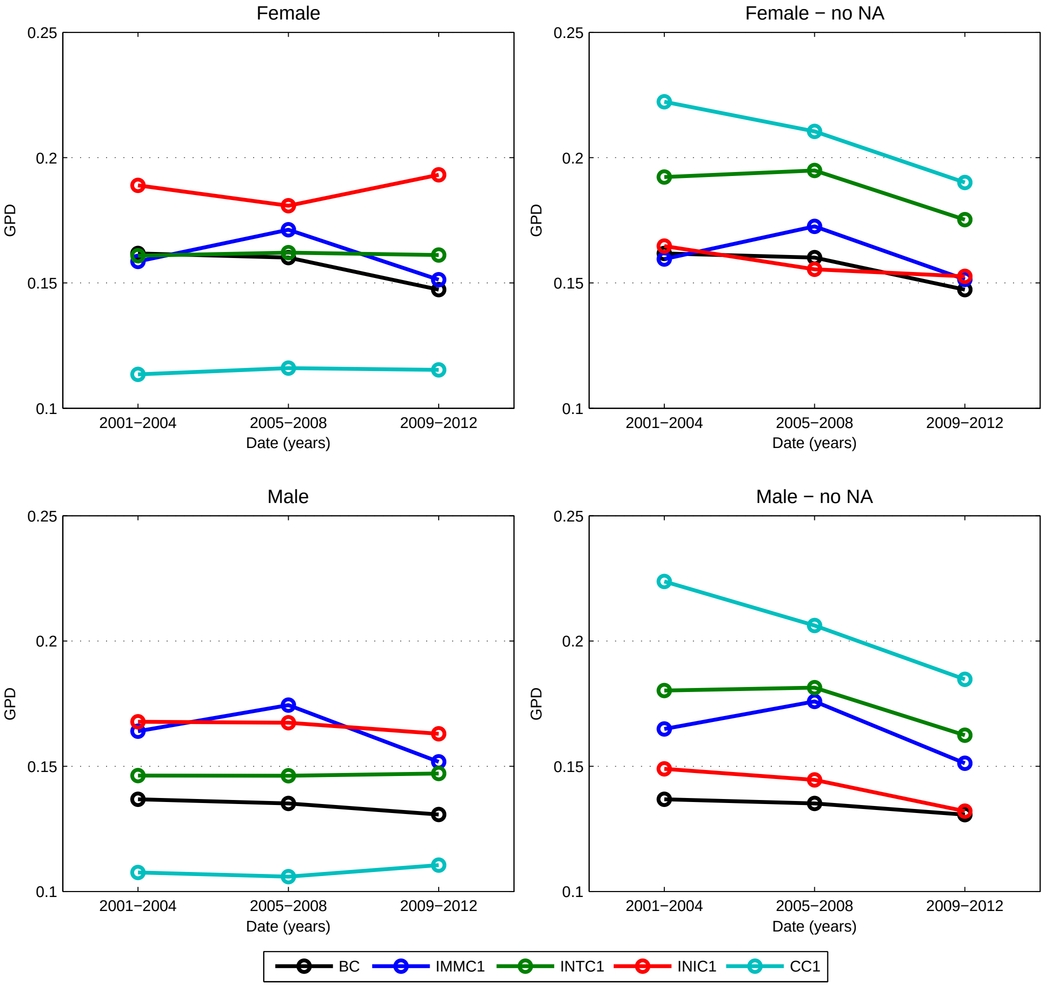


Figure 6.8: Evolution of the Global Probabilistic Deviation (GPD) of *BasicCause* (BC), *ImmediateCause1* (IMMC1), *IntermediateCause1* (INTC1), *InitialCause1* (INIC1) and *ContributiveCause1* (CC1)

to misunderstand 'Antecedent cause´ as clinical antecedents; e.g., the renaming led to the introduction of two prevalent chronic diseases such as hypertensive diseases and diabetes mellitus as antecedent causes, whereas introducing them as contributive causes of death would have been more appropriate. The Spanish National Statistics Institute warned the national Public Health institutions about this problem in 2011. To correct the situation, the term 'Intermediate Cause´ was re-introduced. However, as seen in the results for *IntermediateCause1*, the practice was not abandoned entirely. Finally, the several changes in multiple causes in 2009 carried the problem to the basic cause, which was coded based on the multiple causes. Despite being corrected retrospectively, a small temporal change can still be observed for 2009 (Section 5.3.2, Appendix E). The three versions of the certificates are shown in Appendix E.

Regarding to the detected gradual change, it was probably due to gradual changes in the environment, well represented, e.g., by the increase in life expectancy. Section E.5 in Appendix E shows the temporal heat maps of Age, where this change can be seen.

Regarding to the largest SPO found in Requena (Figure 6.7 (a)), this may reflect an isolated practice in a small department composed of an older population. The increase of SPOs in Torrevieja and Orihuela (Figure 6.7 (b)) may be due to the large number of deaths of young men in Torrevieja. Torrevieja, along the southern coast of the region, has large settlements of immigrants from Eastern Europe and Russia. Other studies have noted the much greater incidence of cancer in Torrevieja and other places close to it probably related to immigration (Zurriaga et al, 2008). Lastly, the dissimilarity between Valencia and other departments, seen in the MSV plot, is mainly due to its lowest proportion of deaths of men in Valencia. Besides, the subgroup of Departments found in the Province of Alicante indicates a local variation of such departments both in terms of the number of filled causes and the causes of death, which may reflect an isolated practice in death certification (for example, we found that 27% of the records were left unfilled with respect to *InitialCause1* in the subgroup of the province of Alicante whereas for the rest, the proportion was 12%).

Finally, despite the GPD metric gradually improved during the period of study (Figure 6.8), for a proper reuse of the registry, users should consider the problems the above-mentioned findings related to variability may cause.

The proposed methods may be adopted in controlling data variability in Public Health research projects or multi-site data-sharing infrastructure. Ensuring DQ requires specific areas of research and investment in public health (Bray and Parkin, 2009; Chen et al, 2014). For example, the WHO recommends conducting regular checks to validate death certification in hospitals as well as investigating new technologies to understand large data sets (WHO, 2012), where the multi-source and temporal methods presented here may prove particularly useful.

6.3 Other case studies

6.3.1 Cancer Registry of the Region of Valencia

Cancer registries are the source of information for national and international cancer statistics as well as for epidemiological research and monitoring of healthcare cancer policies. This case study consists in the application of the multi-source and temporal variability assessment methods developed in this thesis to the Cancer Registry of the Region of Valencia (CRRV), Spain. The CRRV is an Automated Cancer Registry, i.e., patient registries are automatically included from the original samples sent from the different Departments/Hospitals in the Region. This implies that an automated parsing and tumour codification process is done. Hence, with quality control purposes, in automated cancer registries some individual records are manually validated to ensure their correctness with respect to their original cases (Bray and Parkin, 2009; Navarro et al, 2013).

The objective of this case study was two-fold: first, to evaluate the use of the multi-source and temporal variability methods in a real case study, and second, to evaluate the effect that manual case validation could cause to data.

Materials

The CRRV consists of the registered malignant neoplasms in the Region of Valencia between years 2004 and 2013, consisting of a total of 224,267 registries, distributed in 24 Health Departments. The studied variables include demographic data (sex, age, place of birth, place of residence, province of residence, and health department), generic and specific tumour groups, diagnostic base (origin of the diagnosis), validation state (validation state for the registry), and additional disease information (whether the tumour is metastatic, and vital state of the patient). Any information that may disclose the identity of patients was anonymized.

Results

The main findings of this case study are listed as follows:

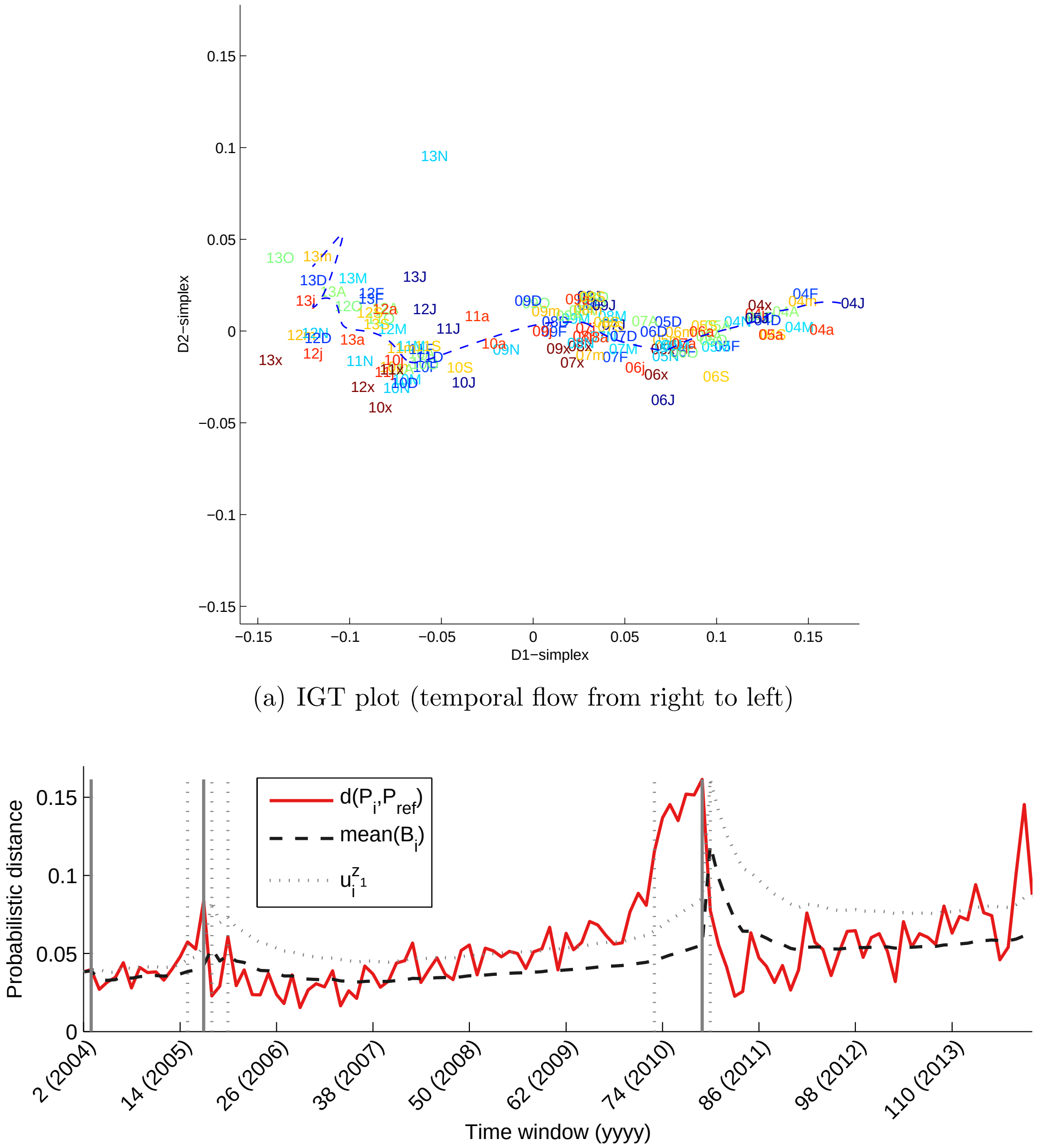
Finding 1. Partitioning into temporal subgroups: The IGT plot and PDF-SPC of the dimensionally reduced registry showed a main abrupt change causing a partition into two temporal subgroups divided around 2010 (Figure 6.9). Additionally, the first of those temporal subgroups also shows minor temporal abrupt changes in 2005 and 2007. Figure 6.10 shows the PDF temporal heat map of the dimensionally reduced registry, on which the changes mentioned above can be found, in addition to the existence of two data clusters through the period of study. At univariate level this change is found as well in the variables related to place of residence, province of residence, health department, place of birth, diagnostic base and validation state.

Finding 2. Temporal changes in generic tumour group and specific tumour group: On these variables the temporal variability methods showed several events during the period of study, specially as slightly isolated temporal subgroups in 2007 and 2010. On these dates, changes in specific tumour incidences were found for 'Primary Unknown' tumours, correlated with opposite changes in others such as 'Lymphomas' and 'Myelomas'. These two variables are the main indicators of incidences of cancer, hence, these changes must be managed carefully.

Finding 3. Recursive conceptual subgroup in validation state in 2007 and 2010: Two isolated conceptual subgroups were found at these dates with a similar pattern with higher frequency of 'Revised' cases and less 'Possible'.

Finding 4. Isolated subgroup of Departments 'La Plana', 'Vinarós' and 'Castellón' in the variables diagnostic base and tumour state during all the period of study: On this variables, the temporal monitoring of multi-source variability methods highlighted a large probabilistic separability of a subgroup formed by these Departments respect to the rest on these variables.

Besides the listed findings, a gradual change is found underlying all the variables what may be expected to some degree due to populational and practice changes. As an example, in this study we can observe a slightly continuous increase in the age of



(a) IGT plot (temporal flow from right to left)

(b) PDF-SPC chart

Figure 6.9: IGT plot (a) and PDF-SPC chart (b) of the dimensionally reduced Cancer Registry of the Region of Valencia (using the first PCA component). A gradual change through the period of study, minor abrupt changes in 2005 and 2007, and a major in 2010 can be observed in the the IGT plot, confirmed by the PDF-SPC, and justified by the data in the heat map in Figure 6.10.

patients, what may be due to the population ageing. Another clear gradual change is found in the metastasis variable, associated to a gradual decrease of metastatic cases found through the period of study. Additionally, we found abrupt decreases of the number of registries in some Departments which remain until the end of the period of study. This was mainly caused to the delay in compiling and sending the sample of cases of the different Departments/Hospitals to the central Public Health Service managing the Cancer Registry.

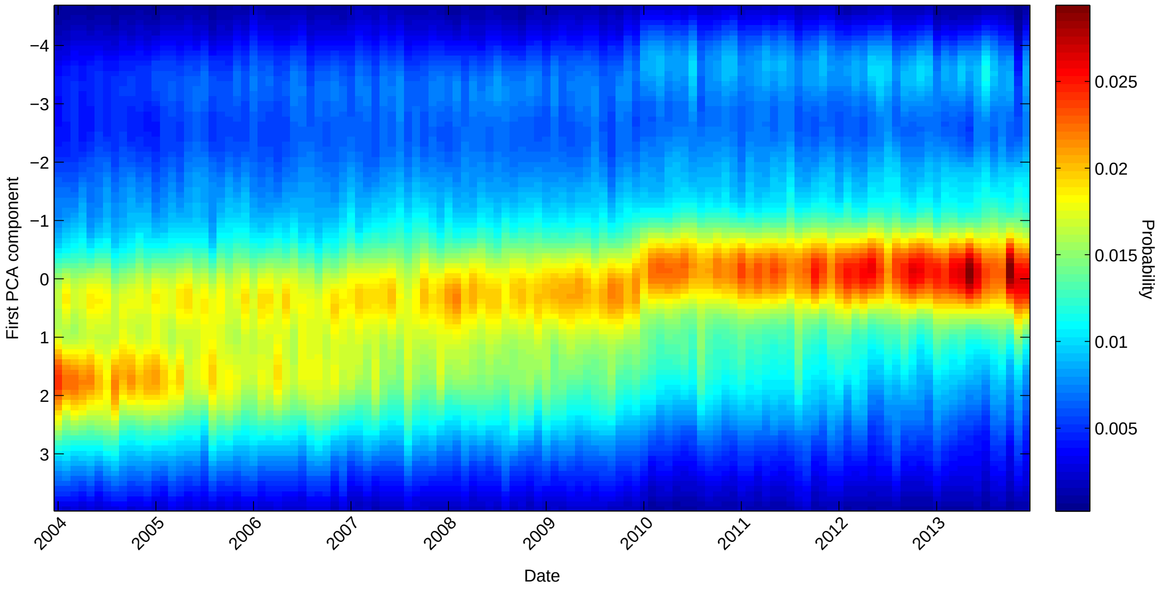


Figure 6.10: Probability distribution heat map (c) of the dimensionally reduced Cancer Registry of the Region of Valencia (using the first PCA component). First, a gradual change through the period of study is observed as a general displacement of the probability mass to the upper part. Next, in 2005 and 2007 minor abrupt shifts of probability masses are observed. Finally, in 2010 an abrupt change is found, mainly observed as a displacement of the probability masses to above. Additionally, the heat map suggests that two clusters of data exist through the period of study as two modes in the dimensionally reduced variable. These are initially centred in 0.25 and 1.75, and affected by the mentioned changes they end centred in 0 and −3.5.

Discussion

After reviewing the findings of this study with the members of the Public Health Service of the Region of Valencia, we found that the four findings are likely related to the effect of case validation.

Hence, we found that the Departments involved in Finding 4 are those which, with a constant quality control, establish a gold-standard subset of the CRRV. Besides, specific quality control actuations were performed in years 2005 and 2010, directly associated to Finding 3. Due to this case revisions, the specificity of the diagnosis tend to increase, what is reflected in Finding 2, as the decrease of the automatically coded as unknown tumours, and increase of specific groups. However, this imply that differences in tumour incidences through time may not be confident in non gold-standard cross-sectional cancer registries. Figure 6.11 shows the effect that these specific case validation actuations had in some generic tumour groups in the evaluated Cancer Registry. Finally, the multivariate contribution of all these findings, together with the decrease of cases due to the delay of sample delivering, led to the temporal subgroups of Finding 1.

The aforementioned effect may have two main consequences. First, time series of automated cancer registries should be interpreted with caution, e.g., breast cancer did not truly increased 1.5 points from 2009 to 2010. And second, automated cancer registries may lead to biased hypotheses or statistical models if these artefacts are not considered. We also remark that specific screening programmes may have similar consequences: increasing the incidence rates of the screened tumour groups.

We can conclude that an external validation with the applied multi-source and temporal methods developed in this thesis may help detecting biases in the data sources of automated cancer registries as well as help measuring the effect of different automated codification procedures.

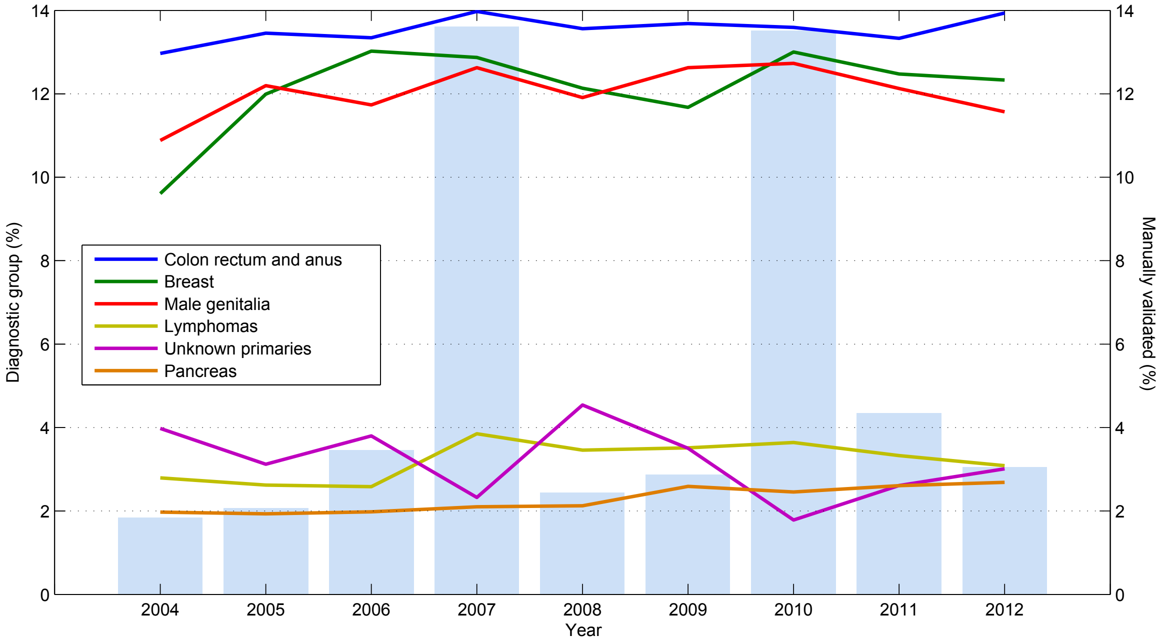


Figure 6.11: Effect of case validation (bars) in the incidences of several cancer diagnostic groups (series) in the Cancer Registry of the Region of Valencia (males and females).

6.3.2 Breast Cancer multi-source dataset

We applied part of the methods developed in this thesis to the data preparation for a predictive model for metastatic affectation in non-sentinel axillary nodes based on the biopsy molecular profile and the sentinel node status in breast cancer.

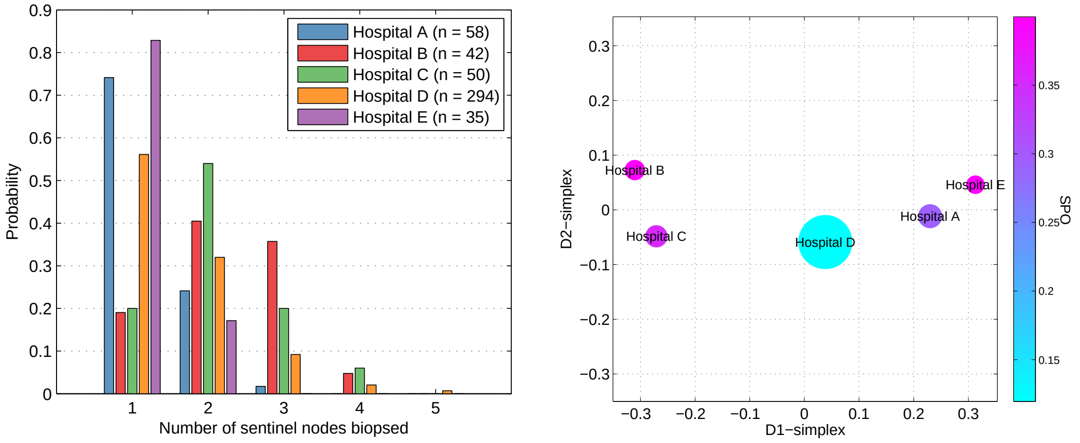
For this case study we applied the multi-source variability method. In addition, we performed a basic DQ assessment, based on the complementary tools developed in this thesis (see Chapter 7), which automatically generated a DQ report informing about missing data, outliers, and the predictive value of the different variables.

Materials

A National Breast Cancer dataset was used, consisting of a sample of 479 cases, with 15 variables related to the biopsy molecular profile and the sentinel node status, distributed in five Spanish Hospitals.

Results

The most relevant finding was a remarkable variability among the Hospitals regarding to the number of biopsied sentinel nodes. This was found by the GPD metrics and MSV plots (Figure 6.12). This finding showed that different protocols were used among the involved Hospitals.



(a) Probability histograms by source.

(b) Multi-source simplicial projection. The color indicates the source SPO.

Figure 6.12: Results of the multi-source variability assessment for the variable *number of biopsed nodes* in the Breast Cancer dataset (hospital names are anonymized).

Discussion

The finding above may lead to possible drawbacks when building a common predictive model using data from all the hospitals. Given to the fact that the target predictive model should be based on the biopsy molecular profile and status, the number of biopsied sentinel nodes would not be used as a independent variable. However, based on the finding of the multi-source variability analysis, such variable was introduced as a corrective factor—associated to the source Hospital—and the accuracy of the models increased, as shown in the work by Bernet et al (2015).

6.3.3 In-vitro Fertilization dataset

The author participated in a project to develop a predictive model for the risk assessment of twin pregnancy in an oocyte donation programme in a private IVF clinic. In this project we followed a customized version of the CRISP-DM data mining methodology (Shearer, 2000), on which first stages data is analysed for their understanding and next preparation for being mined. It is at these stages where we applied the initial versions of the methods of this thesis towards a simultaneous data quality assessment and understanding, facilitating the preparation of a quality assured dataset for the model construction. Hence, we applied the initially developed complementary DQ assessment tools, which automatically generated a DQ report.

Materials

The dataset used in this case consisted of 13,386 In-Vitro Fertilization (IVF) cycles acquired from 2007 to 2010, with a total of 55 variables including clinical and treatment variables from the recipient and the donor, and a set of laboratory variables including oocyte, sperm, and embryo features.

Results

The generated DQ report informed about variable distributions and types, missing data, outliers, the predictive value of the different variables, facilitating the data preparation for the modelling stage. Besides, we provided the first approach of the temporal variability analysis, using the PDF temporal heat maps described in chapter 5. Figure 6.13 shows a sample of the report.

Discussion

The generated DQ report was delivered to the IVF clinic as an official project deliverable. The variables and records which did not achieve a sufficient quality were discarded for the modelling. The DQ procedure helped as well in the feature selection. The final risk assessment model was based on two bayesian logistic regression models. The first provides the probability of an ongoing pregnancy given that one embryo was transferred. The second provides the probability of an ongoing twin pregnancy, given that two embryos were transferred. The outcomes of the two models support the decision of transferring one or two embryos. The model is currently under a prospective validation in the IVF clinic, having reduced the twin-pregnancy rate in a 21% withoug compromising the number of successful pregnancies.

6.4 Limitations

Due to the high number of possible combinations of variables in most case studies, the efficiency of the approach may be improved through automated procedures or a guided Graphical User Interface. Besides, although the methods permit quantitative and qualitative descriptions of variability, it is the duty of the investigator to look for external original causes of variability, based on the insights provided by these methods.

In the Mortality Registry case study we used the PCA dimensionality reduction method because it was simple and enabled us to find the most relevant problems. Although the methods permit analyzing multivariate joint distributions, aiming to simplification and to the reduction of the probabilistic space, other non-linear methods may be better suited to multi-type and multi-modal data. We also found that the PDF-SPC algorithm may require a calibration of its thresholds in some situations to better detect those changes detected by the IGT plot, instead of using the classical three-sigma rule used in the present study. This may related to the latent changes in the distributions due to environmental changes as well as to the width of the analysed time periods, which limit the number of distribution distances accumulated for the monitoring by the PDF-SPC algorithm.

Finally, throughout the case studies, we found that the interpretation of the results of the multi-source and temporal variability methods in some occasions resulted complicated to users. To alleviate this problem we posed an additional effort in facilitating an intuitive description of the methods, as shown in the first section of this chapter, and accompany them by the basic examples shown in Appendix D.

6.5 Conclusions

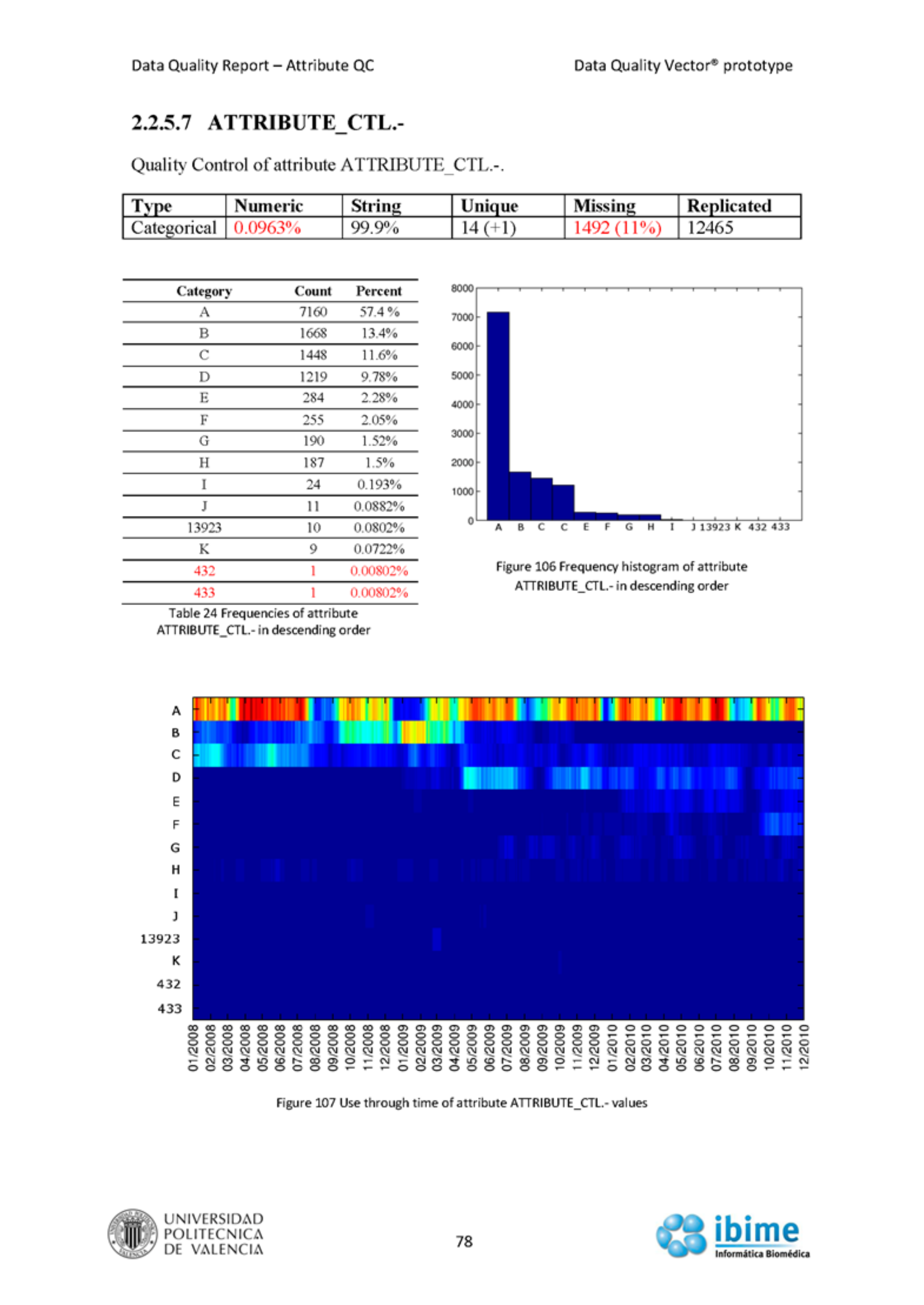
Since the initial evaluation of the first prototypes in the In-vitro Ferlization case study until the validation of the final methods in the Mortality Registry, the methods have shown an evolution being continuously improved during the development of this thesis. The findings in these case studies show that the applied probabilistic methods may result useful as a systematic and generalizable approach to detect and characterize multi-site and temporal variability in multi-site data for their reuse.

Unexpected variability in data distributions among sites or over time can be considered a DQ problem, which may lead to inaccurate or irreproducible results, or suboptimal decisions, when the data are reused. In the In-Vitro Fertilization case study, the methods reduced the time of manually preparing data for their analysis. In the Breast Cancer case study, the methods permitted discovering differences among the involved data sources (hospitals), which may have leaded to a predictive model of poor effectiveness for a global use. In the Cancer Registry, the methods quantitatively confirmed the insights about the effect of manual case revision, in addition to providing a report with a complete description of the variability of the dataset for their better understanding. In the Mortality Registry, the methods mainly evidenced the problem of the change of death certificate and its further amendment, which, being in our case localized to the Region of Valencia, may be translated to National level. Besides, we highlighted differences in the certificate filling practices through Health Departments, what may have consequences in global statistics.

Given all the possible difficulties for a proper data reuse which variability among sources or through time may entail, we suggest that, in addition to integration and semantic aspects, the temporal and multi-site probabilistic variability of data should be incorporated in systematic procedures of assessing DQ to help ensure that valid conclusions are drawn when such data are reused. Additionally, DQ is requiring specific areas of research and investment in Public Health (Bray and Parkin, 2009; Chen et al, 2014). For example, the WHO recommends conducting regular checks to validate death certification in hospitals as well as investigating new technologies to understand large data sets (WHO, 2012), where the multi-source and temporal methods presented in this thesis may prove particularly useful.

Table 6.3: Variability in the Mortality Registry and its Causes. Observable causes are those intrinsic to the data and found during the assessment process. The possible original causes are the external factors that cause the variability. Causes are linked to generic findings in Table 7.1.

Finding (generic code in Table 7.1)	Observable cause	Possible original cause	Detected by
Temporal anomaly in January to March 2000 (F1)	A great deal of missing data in temporal batches	Lack of electronic coding of paper certificate	IGT plot (Figure 6.2 (a)), temporal heat map (section E.3)
Gradual change through the period of study (F2)	Gradual shifts in probability of causes-of-death through time	Increase of life expectancy, social and clinical changes in practice	IGT plots (Figure 6.2 (b), Section 5.3.2), PDF-SPC (Figure 6.3), temporal heat maps (Section E.9)
Abrupt change in March 2009 dividing repository in two temporal subgroups (F3)	Abrupt change in probability of NAs for most variables, and to a small extent in other specific causes of death including the basic cause	Change in the National Certificate of Death	IGT-plot (figure 6.2 (b), Figure 6.4 (a,b)), PDF-SPC (Figure 6.3), temporal heat map (Figure 6.4 (c)), MSV plot monitoring (Figure 6.7)
Other minor abrupt changes in 2005, 2009 and 2011 (F3)	Abrupt changes in probability mass of specific causes of death	National programs for control and prevention of diseases, redesign of certificate, change of disease patterns	IGT plot (Figure 6.4 (a,b)), temporal heat map (Figure 6.4 (c))
Seasonal variations in causes of death (F4)	Seasonality of diseases, mainly winter-specific respiratory diseases and greater incidence of heart diseases in summer	Normal environmental and social effects	IGT-plots (Figure 6.2 (b), Section 5.3.2)
Department of Requena as an outlier (F5)	Requena provides more number of causes, specially contributive causes	Isolated certificate filling practice in the small department with older population	SPO monitoring (Figure 6.6 (a)), MSV plot (Figure 6.6 (b))
Anomalous Department of Torrevieja (F5)	Anomalous population, with more deaths of young men	Different population due to immigration	SPO monitoring (Figure 6.6 (a)), MSV plot (Figure 6.6 (c))
Subgroup composed by Departments in province of Alicante (F6)	More intermediate and initial causes filled but fewer immediate causes. Other differences in incidence of causes.	Isolated certificate filling practices	MSV plot (Figure 6.7)



Data Quality Report – Attribute QC

Data Quality Vector® prototype

2.2.5.7 ATTRIBUTE_CTL.-

Quality Control of attribute ATTRIBUTE_CTL.-.

Type	Numeric	String	Unique	Missing	Replicated
Categorical	0.0963%	99.9%	14 (+1)	1492 (11%)	12465

Category	Count	Percent
A	7160	57.4 %
B	1668	13.4%
C	1448	11.6%
D	1219	9.78%
E	284	2.28%
F	255	2.05%
G	190	1.52%
H	187	1.5%
I	24	0.193%
J	11	0.0882%
13923	10	0.0802%
K	9	0.0722%
432	1	0.00802%
433	1	0.00802%

Table 24 Frequencies of attribute ATTRIBUTE_CTL.- in descending order

Figure 106 Frequency histogram of attribute ATTRIBUTE_CTL.- in descending order

Figure 107 Use through time of attribute ATTRIBUTE_CTL.- values

UNIVERSIDAD POLITECNICA DE VALENCIA

78

ibime Informática Biomédica

Figure 6.13: Page of the DQ report generated for the IVF use case. The basic DQ results and temporal heat map of a variable are shown (variable names and values are anonymized). The categorical variable counts with 0,0963% of numerical values and an 11% of missing data. The most probable value is ‘A’, and values ‘432’ and ‘433’ are marked as possible outliers given their low frequency. Finally, an abrupt change in the probabilities over time is observed in the temporal heat map.

Chapter 7

Biomedical data quality framework

The previous chapters have described the scientific contributions carried out in this thesis. Any scientific work constitutes a step forward, in a minor or major degree, through the global iterative scientific evolution, while contributing to the improvement of population quality of life. And it is at specific points of these continuous scientific evolution when the developed methods and technologies should be transferred to make them applicable to real-world scenarios. This was a consideration taken since the beginning of this thesis for two main reasons. First, for the sensible domain and implications that the quality of data may have in the outcomes of research and healthcare. And the second is a practical reason: given the complexity of the experiments to be carried out, on massive data from several case studies, several variable and distribution types, and with landmarked batch analyses, a software toolbox which eases the repetition with different settings would improve the efficiency of the research and its translation to practice.

This chapter is divided in three sections. First, the proposed systematic approach and developed software for multi-source and temporal variability assessment is described. Second, a definition of a theoretical basis for a general framework for the evaluation of DQ in biomedical data is described. This framework includes the multi-source and temporal variability DQ aspects. Finally, three applications of this theoretical DQ framework are discussed: in a process for the construction of quality assured infant feeding repositories, for the contextualization of data for their reuse in CDSS, and in a on-line service for the evaluation and rating of biomedical data repositories.

The systematic approach described in this chapter was published in the journal publication by Sáez et al (2016)—thesis contribution P5. Parts of the description of the general DQ framework were published in the conference paper by Sáez et al (2012b)—thesis contribution P1. The developed software toolbox corresponds to the software contribution S1, and is registered in the technological offer of the UPV. The derived application of the perinatal quality assured repositories is under review as two publications in the Computer Methods and Programs in Biomedicine journal. The derived application of the HL7-CDA wrapper for data contextualization in CDSSs was published in the journal publication by Sáez et al (2013a), selected by the IMIA as one of the best medical informatics papers published in 2013 in the subfield of Health Information Systems—thesis publication P6 and software contribution S2. The derived *appliation of the construction of quality assured perinatal repositories has been accepted as journal publication by García de León Chocano et al (2015)—thesis publications P8 and P9. The derived application of Qualize corresponds to software contribution S3.*

7.1 Multi-source and temporal variability

This section is divided in two parts. First, we describe the systematic approach for the assessment of multi-source and temporal variability using the methods developed in this thesis, which is proposed based on the experience of their application in the case-studies. Second, we describe the software for DQ assessment developed in this thesis, which stands as the software solution to be used in the proposed systematic approach.

7.1.1 Systematic approach

The case studies presented in the previous chapter allowed the validation of the methods developed during this thesis research into real-world problems. For the case studies on which we analysed the multi-source and temporal variability we counted with their respective assessment metrics and visualizations, namely the GPD, SPO, MSV plot, IGT plot, PDF-SPC and temporal heat maps. As a consequence, the discovered findings were reached within a systematic use of these methods.

Based on the experiences of applying these methods in the case studies, we propose a systematic approach to assess the multi-source and temporal variability in repositories of biomedical data (Figure 7.1). In a top-down approach, one starts by analyzing the temporal or multi-site variability of the complete data set and then, based on the results of the analysis and prior knowledge of the repository, drills down to specific variables or groups of variables. The process can be cyclic, similar to an On-Line Analytical Processing (OLAP) exploratory analysis, navigating through different levels of granularity; for example, a temporal change found in the complete repository could be caused by a sudden bias within a single site. Such an anomalous site may require a specific temporal analysis, and excluding it may facilitate the discovery of other patterns or sources of variability.

The proposed generalizable approach may be adopted in controlling data variability in research projects or multi-site data-sharing infrastructure. Hence, this approach can help discovering different findings related to the variability in data distributions which may require different solutions for a proper data reuse. A selection of them is described in Table 7.1, in which we attempt to provide a generic list of findings related to multi-source or temporal variability in repositories of biomedical data along with their possible causes, problems in reusing the data, and solutions.

The problems listed in Table 7.1 are associated with basic research uses of data, e.g., for empirical derivation of hypotheses or statistical models. The proposed solutions vary with the sites or time affected and include fixing or excluding data or analyzing distinct groups of sites or time periods separately. For example, for statistical modeling, an abrupt temporal change may reduce the model's effectiveness when using the

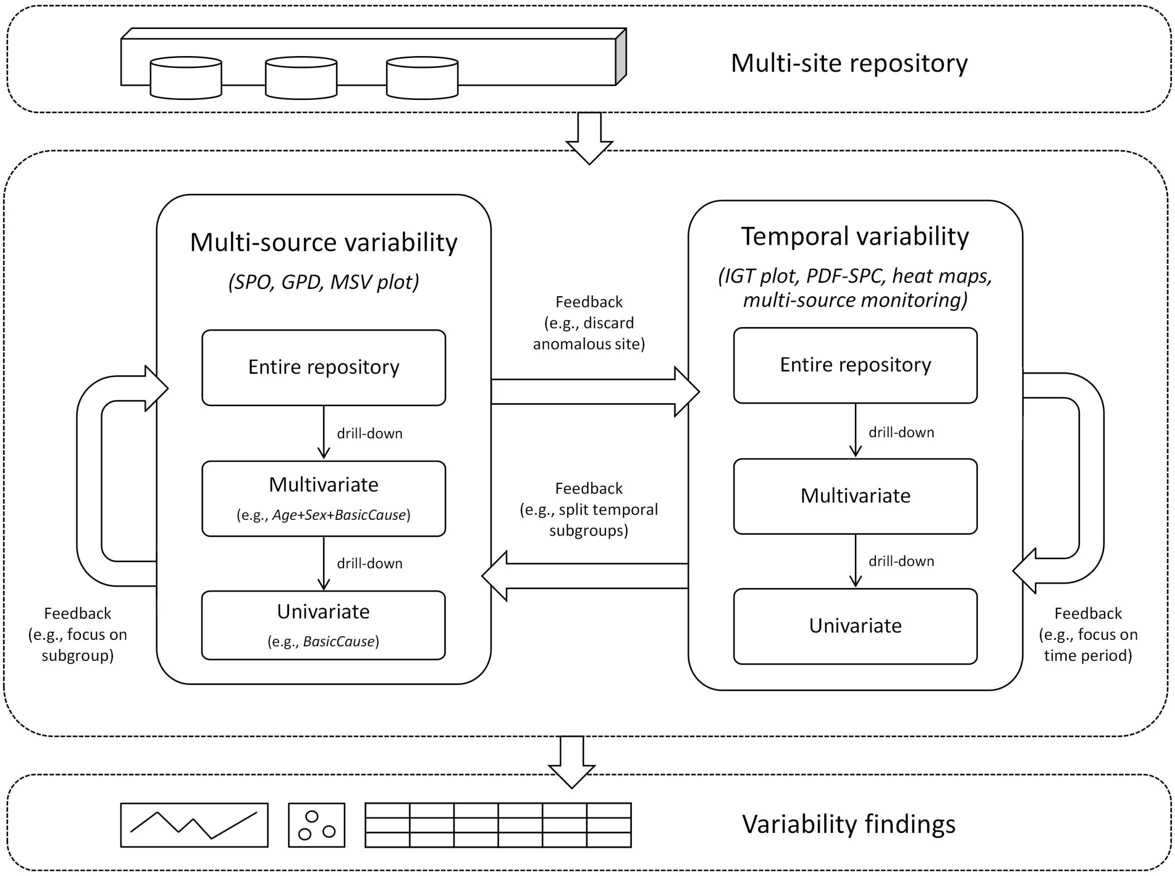


Figure 7.1: Proposed Systematic Approach to Assess the Temporal and Multi-Site Variability of Repositories of Biomedical Data Using Probabilistic Data Quality Control Methods.

data for the entire period: if the change is due to an environmental change (e.g., a change of protocol), and not to any error, separate models for periods before and after the change would yield better results, and a model with a good further generalization would be one giving more importance to latest data. Besides, a probabilistically isolated site or group of sites may bias the results of a global analysis—as illustrated by García-Gómez et al (2009) in the case of multi-site predictive models for brain tumor diagnosis. Excluding biased sites would improve the global results and in the case of multi-site subgroups, a good solution would be to analyze them separately. An alternative solution which may reduce user involvement could be using incremental learning approaches, which rank the data in terms of importance by their age (Gama and Gaber, 2007) or provenance (Tortajada et al, 2011). Fixing problematic data may also be considered when variability is associated with intrinsic problems with DQ such as changes in the degree of data completeness or consistency. It is important to note that variability among sites or with time does not imply an error but requires the lack of concordance to be investigated. In fact, in some cases, variability may be inherent in biomedical data because the data are affected by the environment, population, or other external factors such as a programmed change in protocol. However, in other cases, variability may be unexpected, e.g., that due to faulty acquisition processes or biased actuations, which could include biased or faulty data input, system design, or variation in healthcare quality.

As an example to facilitate the identification of these generic findings as specific findings, we have matched the generic findings in Table 7.1 with those found in the Mortality Registry case study, as shown in Table 6.3.

7.1.2 Developed software toolbox

The development of the research and experimentation carried out in this thesis supposed a technological challenge. Two were the main reasons. First, the proposed methods were based on several complex methods for multiple objectives:

- estimating probability distributions,
- dealing with different types of uni and multi-variate variables,
- calculating PDF distances,
- estimating simplicial projections and statistical manifolds,
- incrementally estimating PDFs on temporal landmarks,
- smoothing temporal data,
- plotting results based on adequate visual analytics, and
- automatically generating reports.

And second, with the purpose to carry out the several experimentations and case studies, the methods should be used in a systematic, configurable and replicable manner, and if possible, with an efficient computational cost.

These challenges were considered since the beginning of the thesis development. Step by step, a generic and systematic software toolbox (developed in MATLAB®) for multi-source and temporal variability assessment was constructed. From its initial until the latest versions, the framework was applied to different real case studies (Chapter 6). In addition to the methods for multi-source and temporal variability, basic DQ profiling methods for completeness, consistency and uniqueness dimensions were included to facilitate the DQ assessment. This software toolbox can be used as the basis for an industrialization of the multi-source and temporal variability assessment methods.

The developed software toolbox is divided in six main packages, described next and summarized in Figure 7.2.

Probabilistic framework: Contains all the functions related to the probabilistic framework for both multi-source and temporal variability. This includes classes for representing and estimating PDFs for continuous and categorical data, using MATLAB object polymorphism. While the *histogram* class is used to represent the PDF of a data sample, which can be of different variable types, the class *temporalHistogram* contains the several single histograms of the consecutive temporal batches (see 'Supporting

PDF classes' in Figure 7.2). The PDFs are estimated from input data files based on an incremental/streaming manner, what permits estimating the PDFs of Big Data files which cannot be entirely loaded in the computer memory. In addition, the estimation functions can be carried out under a parallel computing approach, what was tested in the Distributed Computing Server of the ITACA Institute at UPV, which counts with a distributed MATLAB server on seven rack computers with a total of 84 processor cores, 168 threads and 64GB of memory. This package included additional tools such as PDF distance calculations, date operation functions, and automatic variable type inference.

Multi-source variability: Contains the methods for the multi-source variability assessment (described in Chapter 4). Hence, it provides a function which receives a set of PDFs from different sources and returns the GPD and SPO metrics and the coordinates and centroid of the resultant simplicial projection. It also provides access to this metrics using the data itself, with internal estimation of PDFs for continuous or categorical data. Finally, it provides the corresponding functions for generating the output visualizations, including the 2D and 3D MSV plots and simplices, phylogenetic trees, and basic PDF histograms and densities comparisons for the multiple sources.

Temporal variability: Contains the methods for the temporal variability assessment (described in chapter 5). Hence, it provides a method which takes a univariate or multivarate dataset and, given a temporal landmark (e.g., days, weeks or months), computes the statistical manifold for the IGT plot and performs the PDF-SPC monitoring. It additionally includes PDF smoothing functions to be used in the previous functions, which smooth temporal PDFs based on sliding windows, fading windows, memory-less fading windows and temporal landmarks. Finally, it provides the corresponding functions for generating the output visualizations, including the IGT plot, PDF-SPC control chart, and absolute frequencies and PDF temporal heat maps.

Multi-source monitoring: Contains the methods for the monitoring over time of the multi-source variability metrics and visualizations. Concretely, permits calculating the GPDs, SPOs and simplices from multi-source data over a time period, as well as plotting their results and generating a video of the evolution of data sources in the simplicial probabilistic space.

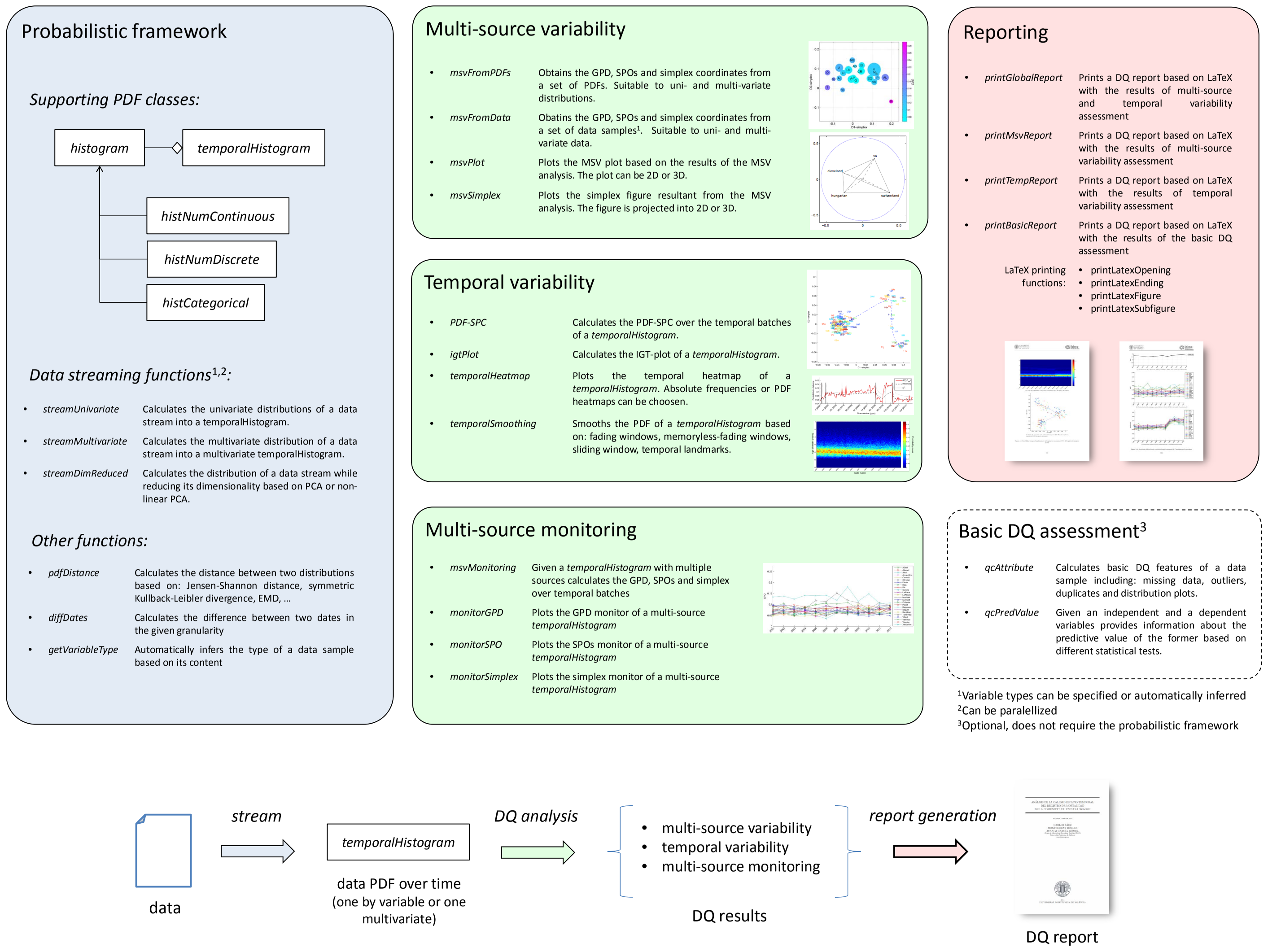
Reporting: This package contains the required functions for the automatic report generation facilitated by the framework. These functions take the results and output figures of the previous modules and dynamically creates LaTeX[e] code which is then compiled into a .pdf file—do not confuse with the Probability Distribution Function acronym—file. Hence, it contains a global report function, supported by specific functions for writing descriptive results, and those of multi-source, temporal and spatio-temporal analysis.

[e] LaTeX document preparation system http://latex-project.org/ (accessed 2015-09-24)

Basic DQ assessment: This package includes the initial methods developed for the assessment of the basic DQ features of missing data, outliers, duplicates and simple plots. This cannot be considered part of the multi-source and temporal variability framework, as it does not make use of the common probabilistic framework or methods, but their results can be as well printed in a basic DQ report.

Table 7.1: Generic temporal and multi-site variability findings and possible causes, problems and solutions. The causes are linked to findings in the Mortality Registry case study in Table 6.3.

Generic Finding (code)	Detection method	Generic Possible Cause	Possible Data Reuse Problems	Possible Solutions
Punctual temporal anomaly (F1)	IGT plot, heat maps	Biased temporal batch	Biased container time period (a year given a biased month), inaccurate research hypotheses or statistical models	Fix temporal batch; remove container time period
Gradual change (F2)	IGT plot, PDF-SPC, heat maps	Normal evolution of population or clinical practice	Outdated statistical models	Incremental learning of models
Abrupt change causing temporal subgroups (F3)	IGT plot, PDF-SPC, heat maps	Change of protocols, systematic errors, environmental or social effects	Inaccurate research hypotheses or statistical models: results that are not concordant before or after	Separate analyses, incremental learning of models
Seasonality (F4)	IGT plot	Normal environmental or social effects	Inaccurate statistical models	Season-specific models
Anomalous sites (F5)	MSV plot, SPO, GPD	Anomalous population, biased clinical practice or systematic errors	Biased research hypotheses or statistical models: incompatible decisions or models among sources	Separate analyses or separate models for outlying sites
Multi-site subgroups (F6)	MSV plot, SPO, GPD	Groups of sources with isolated populations, clinical practices or systematic errors		Separate analyses or separate models for subgroups



Probabilistic framework

Supporting PDF classes:

Data streaming functions[1,2]*:*

- *streamUnivariate* Calculates the univariate distributions of a data stream into a temporalHistogram.
- *streamMultivariate* Calculates the multivariate distribution of a data stream into a multivariate temporalHistogram.
- *streamDimReduced* Calculates the distribution of a data stream while reducing its dimensionality based on PCA or non-linear PCA.

Other functions:

- *pdfDistance* Calculates the distance between two distributions based on: Jensen-Shannon distance, symmetric Kullback-Leibler divergence, EMD, ...
- *diffDates* Calculates the difference between two dates in the given granularity
- *getVariableType* Automatically infers the type of a data sample based on its content

Multi-source variability

- *msvFromPDFs* Obtains the GPD, SPOs and simplex coordinates from a set of PDFs. Suitable to uni- and multi-variate distributions.
- *msvFromData* Obatins the GPD, SPOs and simplex coordinates from a set of data samples[1]. Suitable to uni- and multi-variate data.
- *msvPlot* Plots the MSV plot based on the results of the MSV analysis. The plot can be 2D or 3D.
- *msvSimplex* Plots the simplex figure resultant from the MSV analysis. The figure is projected into 2D or 3D.

Temporal variability

- *PDF-SPC* Calculates the PDF-SPC over the temporal batches of a *temporalHistogram*.
- *igtPlot* Calculates the IGT-plot of a *temporalHistogram*.
- *temporalHeatmap* Plots the temporal heatmap of a *temporalHistogram*. Absolute frequencies or PDF heatmaps can be choosen.
- *temporalSmoothing* Smooths the PDF of a *temporalHistogram* based on: fading windows, memoryless-fading windows, sliding window, temporal landmarks.

Multi-source monitoring

- *msvMonitoring* Given a *temporalHistogram* with multiple sources calculates the GPD, SPOs and simplex over temporal batches
- *monitorGPD* Plots the GPD monitor of a multi-source *temporalHistogram*
- *monitorSPO* Plots the SPOs monitor of a multi-source *temporalHistogram*
- *monitorSimplex* Plots the simplex monitor of a multi-source *temporalHistogram*

Reporting

- *printGlobalReport* Prints a DQ report based on LaTeX with the results of multi-source and temporal variability assessment
- *printMsvReport* Prints a DQ report based on LaTeX with the results of multi-source variability assessment
- *printTempReport* Prints a DQ report based on LaTeX with the results of temporal variability assessment
- *printBasicReport* Prints a DQ report based on LaTeX with the results of the basic DQ assessment

LaTeX printing functions:

- printLatexOpening
- printLatexEnding
- printLatexFigure
- printLatexSubfigure

Basic DQ assessment[3]

- *qcAttribute* Calculates basic DQ features of a data sample including: missing data, outliers, duplicates and distribution plots.
- *qcPredValue* Given an independent and a dependent variables provides information about the predictive value of the former based on different statistical tests.

[1] Variable types can be specified or automatically inferred
[2] Can be paralellized
[3] Optional, does not require the probabilistic framework

Figure 7.2: Modules and main functions of the developed software toolbox and example of use.

It is worth to mention one difficulty that was overcome in the streaming estimation of the PDFs in temporal batches. First, the temporal analysis is made based on analysing data through temporal batches at a given time period, such as weeks or months. Therefore, in some case studies we found that, due to the frequency at which cases were acquired, some time periods counted with no data filling specific ranges in the variable support. As a consequence, some intermediate PDFs counted with bins with a probability of 0, what impeded properly calculating PDF distances. This issue was solved using static PDF smoothing mechanisms, such as absolute discounting for categorical data (Ney et al, 1994), KDE for continuous, or the temporal smoothing methods mentioned above. A more important problem occurred when no individuals were observed in a PDF at a given period, leading to a probability which sums 0 what, in fact, by definition is not possible. Consequently, as a convention we assumed that any distance to 0-probability PDFs would be the maximum possible, what permitted using normally the developed methods on these situations.

Finally, it is also worth to mention that most of the figures in this thesis showing results of the multi-source and temporal variability were obtained using this framework.

7.2 Towards a general data quality framework

The quality of data is of great importance for a valid and reliable data reuse. As described in Section 2.1, many studies aim to provide methods and dimensions to assess the quality of biomedical data. However, depending on the type of data reuse, what it is defined as quality may vary. Consequently, any guideline or framework facilitating what and how to assess data quality, independent of the type of data reuse, would help in the definition and establishment of data quality assurance protocols for biomedical data reuse.

In this section we describe the aspects related to the definition of a theoretical framework for the evaluation of DQ in biomedical data repositories. This framework is based on the definition of nine DQ dimensions, including the multi-source and temporal methods and dimensions, aiming to cover the most important aspects to our knowledge in the literature. Dimensions can be measured in different axes of the dataset, namely through registries, attributes, single-values, entire datasets, multi-source and through time. Several examples for these measurement possibilities are discussed next.

The objective of this proposal is to provide insights into further research in other DQ dimensions alone or in combination with the multi-source and temporal variability problems, towards the application and industrialization of a general DQ framework.

7.2.1 Functionalities and outcomes

In Section 2.1 we reviewed the biomedical data quality state-of-the-art, and introduced some of the origins of the general discipline. Recent published reviews (Batini et al, 2009; Weiskopf and Weng, 2013; Liaw et al, 2013) plus ours, found that there is little agreement in the data quality methods and dimensions being addressed. Scientific papers usually suggest either global data quality frameworks (Wang and Strong, 1996; Jeusfeld et al, 1998; Lee et al, 2002; Pipino et al, 2002; Arts et al, 2002; Pierce, 2004;

Oliveira et al, 2005; Karr et al, 2006; Choquet et al, 2010) with a selection of dimensions and methods, or concrete solutions for specific data quality problems (Choi et al, 2008; Heinrich et al, 2009; Etcheverry et al, 2010; Weiskopf and Weng, 2013; Alpar and Winkelsträter, 2014). On the other hand, commercial solutions for data quality assessment usually provide general purpose data profiling methods or rule-based data quality checks (Judah, Saul and Friedman, Ted, 2014). In our case, in contrast, we aim to provide a closer focus to the assessment of biomedical data repositories.

The first step for defining the general expected features of our DQ assessment framework was defining the system's expected functionalities and outcomes. Table 7.2 classifies the expected functionalities, while table 7.3 classifies the expected outcomes of the system. With these tables, we intend to facilitate deciding the objectives of a DQ assessment system or protocol. We remark that the classified functionalities and outcomes can be related among them, inter and intra-table. For instance, a DQ monitoring system can check DQ alerts based on temporal analysis.

In this thesis we mainly focused to the assessment of biomedical data repositories at the population level, i.e. based on the sample distributions. However, the data quality assessment of a single case, as shown in the first row of Table 7.2, is important enough to be considered since the first step towards reducing DQ problems is preventing issues at the acquisition of data individuals.

Table 7.2: Defined functionalities for the DQ assessment of biomedical data

Functionality	Description
Single case quality assessment	DQ analysis for a single case in insert, update or retrieval time
Data repositories quality assessment	DQ analysis for a complete data repository, generally reuse repositories for research or decision making
Continuous DQ monitoring	A monitor of the DQ of streams or batches
Alerts about DQ	The system triggers an alert based on predefined DQ rules
Selection of quality assured data	The user wants to obtain a set of data that fulfils a set of DQ requirements
DQ reports generation	Obtain a DQ report based in a predefined or custom DQ assessment query
Data integration	Control and assure the DQ in the integration of data in a centralized or federated database

7.2.2 Data

Next, we define the characteristics of data that may be used as input for a DQ assessment analysis. These are divided into how data is accessed, and what types of variables are analysed.

Table 7.3: Outcomes of DQ assessment for biomedical data

Classification	Description
DQ metrics	The measurements of DQ dimensions or functions of them
DQ visualizations	Visualizations for the exploratory analysis and visual inspection of DQ results
Set of high-quality data	A set of data that fulfils some DQ requirements
Track of low DQ causes	Hints for the possible causes of recurrent low DQ
Trends of DQ	An analysis of trends of DQ
DQ report	A document describing DQ the results of analysis and findings

Data access: Regarding to the data accesses, we classified two groups, as shown in Table 7.4: off-line datasets and on-line (or streaming) data analysis. Table shows the results of this classification. This classification is principally suited to the analysis of sets of data or data repositories, not applicable for the single-case assessment. We also classified how these data accesses may affect to the temporal DQ analysis.

Table 7.4: Types of data accesses and relation to the temporal DQ assessment

Data access	Temporal analysis metadata	Temporal assessment possible
Off-line (file dataset, local database)	Non-timestamped	✗
	Timestamped	✓
On-line (streams, batch analysis)	Landmarked batches	✓
	Fixed frequency batch analysis	✓

Regarding to the single-case assessment, the on-line equivalent would be using an automatic DQ control validation at the time data is acquired. E.g., an automatic alerting method may warn about possible inconsistent values being introduced, or once a complete case is registered the system may check for missing important values, multivariate inconsistencies, or whether the case is classified as an outlier. The off-line equivalent would be applying single-case DQ analyses, as those just described, case-by-case in a registry or focusing on a specific case.

Data types: Biomedical data can be seen as a set of registries composed by atomic elements representing real-world entities, either patient observations or contextual information. Table 7.5 shows a proposal for the high-level data types that can be present in a registry to be used in the DQ measurements—not to be confused with their probabilistic variable types.

Registry identifiers may be the first source on which to search for duplicated registries. Besides, being registry identifiers generally unique, there may be not sensible to analyse their distributions in other DQ assessments such as the multi-source and

Table 7.5: High-level data types

Type of element	Examples
Registry identifier	Patient ID, protocol ID
Numerical observation	Age, BMI, blood pressure
Categorical observation	Gender, applied therapy, diagnosis
Complex observation	Image, signal, free text
Contextual	Clinical domain, data source, discharge date

temporal variability. The three types of observations: numerical, categorical and complex, generally represent the patient status within the registry, and their DQ can be analysed based on generic and domain-specific methods. Finally, data typed as contextual would represent the context at which the data in the registry was acquired, including data sources and timestamps.

This classification might result simple, however, we believe it is a high level representation for most situations, which resulted after a review of data types from those proposed in the literature and in data mining software solutions, as shown in Table 7.6. Hence, registry identifiers are generally typeless or categorical variables. Numerical and categorical observations are respectively numerical and categorical variables. Complex observations may be represented as free text, structured data, graphs, or matrices (e.g., a matrix of numerical). Finally, contextual elements may be of any type if they intend to act as accompanying context information.

Note that numerical and categorical types will be those candidates for the analysis by the multi-source and temporal methods developed in this thesis, and using the PDFs as described in Section 2.2.1. Besides, contextual types may provide the required source and temporal metadata. Also note that these are univariate types, except the matrix/complex type, where multivariate may mix several of these.

7.2.3 Data quality dimensions

In Section 2.1.1 we reviewed the concept of DQ dimensions, as the attributes that represent a single aspect or construct of DQ to be addressed. Despite the wide range of approaches, a degree of agreement in high level concepts was observed, as shown in the proposals of dimensions in Tables 2.1, 2.2 and 2.3. As mentioned, we found an insufficient attention to the problems of data variability among sources or through time, having those been addressed in this thesis.

Therefore, we aimed to define the requisites of a data quality assessment framework which gave an special attention to variability problems, while maintained generic enough to other dimensions. For the proposal of this DQ framework, we intended to make a selection of DQ dimensions, aimed to biomedical domain, which take concepts both from the literature review carried out in this thesis but also based on the experience of the authors in biomedical data analysis. Hence, our proposal of DQ dimensions is shown in Table 7.7.

Table 7.6: Review of variable types in literature and data mining software. Note: ‘NA’ means ‘not applicable’ and ‘cat.’ means ‘categories’.

Proposed			**Weka** (Hall et al, 2009)	**RapidMiner** (Hofmann and Klinkenberg, 2013)	**Clementine 12.0 (SPSS©)**	**Stevens (1951)**	**Hastie et al (2009)**	**UNESCO** (Nagpaul, 1999)
Quantitative (numerical)	Continuous	Continuous interval	Real	Real	Range	Interval	Quantitative	Interval
		Continuous Ordinal	Real	Real	Range	Ordinal	Quantitative	Continuous ordinal
		Continuous Ratio	Real	Real	Range	Ratio	Quantitative	Ratio
	Discrete	Discrete Interval	Integer	Integer	Range	Interval	Quantitative	Interval / Ordinal
		Discrete Ratio	Integer	Integer	Range	Ratio	Quantitative	Ratio / Ordinal
Qualitative (categorical)		Binomial	Nominal / Integer	Binominal	Flag	Nominal-categorical / Interval	Qualitative (2 cat.)	Normal
		Categorical	Nominal	Polynominal	Set	Nominal-categorical	Qualitative (> 2 cat.)	Normal
		Categorical Ordinal	Nominal / Integer	Polynominal / Integer	Ordered set	Ordinal	Ordered categorical	Ordinal / Dummy / Preference
Typeless			NA	NA	Typeless	NA	NA	NA
Date			Date	Date-Time	Range	NA	NA	NA
Free text			String	Text	Typeless	NA	NA	NA
Multivalue			Relational	NA	NA	NA	NA	Multiple-response
Matrix			NA	NA	NA	NA	NA	NA
Structure			NA	NA	NA	NA	NA	NA

Table 7.7: Proposal of DQ dimensions to be addressed in the framework

Dimension	Definition
Completeness	Degree to which relevant data is recorded
Consistency	Degree to which data satisfies constraints and rules, including concordance of units, or impossible values or combinations of values
Uniqueness	Degree to which data contains replicated registries or information representing the same entity
Correctness	Degree of accuracy and precision where data is represented with respect to its real-world state
Temporal stability / Timeliness	The degree of changes in the data probability distributions over time or, according to the timeliness dimension in literature, whether registered data is up-to-date
Multi-source stability	Degree to which data probability distributions are concordant among different sources
Contextualization	Degree to which data is correctly/optimally annotated with the context in with it was acquired
Predictive value	Degree to which data contains proper information for specific decision making purposes
Reliability	Degree of reputation of the stakeholders and institutions involved in the data acquisition

The completeness, consistency, correctness and uniqueness dimensions are generally used in the DQ literature. Although sometimes the first three can overlap on their definitions, or be contained within each other, we recommend making them orthogonal. E.g., a patient observation is incomplete if it is not registered, inconsistent if it is outside a range, or incorrect if, even consistent, it is unlikely to be true.

It can be noted that two of the dimensions refer to temporal and multi-source stability. We must highlight that these two are related to the temporal and multi-source variability approaches developed in this thesis. However, given that the rest of dimensions are expressed in positive terms, we were forced to change ‘variability’ to ‘stability’ in turn. As an example, to measure the multi-source stability we inverted the GPD metric in Equation 4.11 as $1 - GPD$.

These definitions of dimensions present some novelties. Timeliness has generally been used for outdated data, but if data is viewed as an evolving stream, analysing its temporal stability as a data stream problem (Sáez et al, 2015) is a novel DQ concept. Similarly, the novel multi-source variability dimension aims to measure the probabilistic concordance of data among different data sources such as hospitals, physicians, devices, etc. (Sáez et al, 2014b). Besides, contextualization of data is associated to the semantic normalization and annotation of EHRs, however it has not been defined as a measurable DQ dimension yet. Annotated data permits not only understanding data, but also interpreting its quality under different contexts, i.e. using context-specific DQ metrics.

With the selection of DQ dimensions in Table 7.7 we try to define the most important aspects to be addressed to our opinion, which could cover the necessary aspects for the DQ assessment for data reuse while being to some degree personalized on their methods. Dimensions can conform to data specifications or to user expectations. In both cases it is recommended defining an information quality specification (English, 2006), which may refer to aspects of data in their own right and to quality requirements for specific contexts of use. Other authors classify them as contextual assessments (Pipino et al, 2002; Shankaranarayanan and Cai, 2006) where different measurements for a single dimension can exist in both groups. Hence, we propose that some of the dimensions could be classified as generic (i.e. domain-independent, such as a degree of the number of duplicated data) and some others as domain dependent (parametrized given a scenario, such as measuring the predictive value for a specific decision support task).

7.2.4 Axes

We define an axis as the target of the DQ analysis across the provided data structure. Assuming data is provided in tabular format, a registry is represented by a row composed by the set of columns associated to the different variables. Hence, in our proposal, similarly to the work by Oliveira et al (2005), DQ can be analysed on this data table over the axes shown in Table 7.8.

For example, methods such as the proposed in this thesis, which aim to samples of several individuals, would apply to attribute or dataset axes, while other methods for other DQ dimensions would analyse value o registry axes, e.g., analysing the consistency among two values of a registry.

Table 7.8: Axes on which to measure the data quality

Axis	Definition
Value	The value of a single variable (single-case, univariate)
Registry	A patient registry composed by several variables (single-case, multivariate)
Attribute	The values of a variable of a sample of registries (univariate, sample)
Dataset	The values of two or more attributes, until the complete dataset (multivariate, sample)
Time	A comparative analysis of data through time at any of the value, registry, attribute or dataset axes
Source	A comparative analysis of data among sources at any of the value, registry, attribute or dataset axes

Figure 7.3 represents the measurement axes over a tabular data representation. The last two axes are related to the temporal and multi-source stability DQ dimensions, which can be defined as ‘axeable’ dimensions. Hence, the other DQ dimensions can also be measured through such temporal and spatial axes. This way, data could be seen as a multi-way matrix. Then, if a DQ assessment procedure receives a tabular dataset as described before, it could extract the temporal and multi-source axes from the contextual elements identifying a data source or timestamp. Thus, the original dataset

could provide either different subsets of data according to its source, or continuous batches of registries according to a temporal window size.

Analysing any dimension in combination with time, leads to the concept of DQ monitoring. Analogously, the combination with multi-source stability leads to DQ source auditory. As a consequence, according to the temporal axis the different metrics for each dimension can be monitored through time by means of metric series and quality control charts. On the other hand, according to the spatial axis DQ metrics can study differences among the DQ of different data sources, e.g., in an institutional quality of healthcare auditory process.

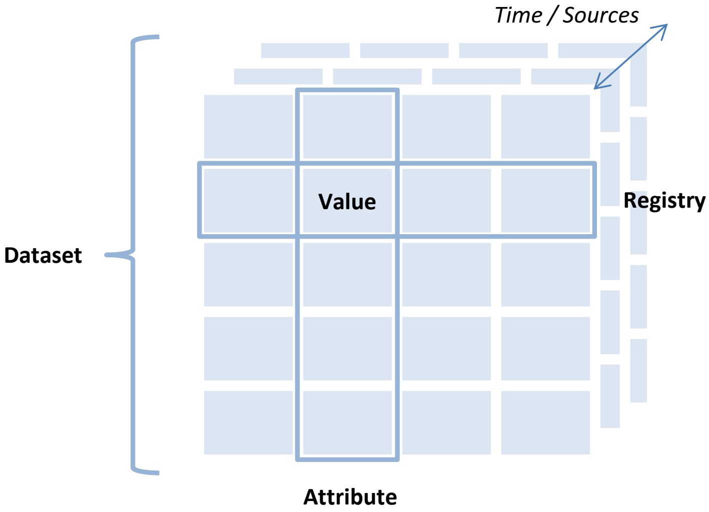


Figure 7.3: Illustration of the data axes at which DQ can be measured in the proposed model

7.2.5 Measurements of (dimension,axis) pairs

The proposed data quality framework offers a model for the definition of DQ metrics based on DQ dimensions and data measurement axes. We propose that DQ dimensions can be measured at the different axes as a (dimension,axis) pair. Each dimension-axis pair can have associated generic or context-specific metrics.

As a consequence, metrics in a dimension can be defined differently according to a target axis. E.g., (completeness,value) measures whether the value of an element is recorded, (completeness,registry) can provide an account of missing data in the registry, (completeness,attribute) can measure the percentage of missing data throughout the attribute, and (completeness,dataset) whether the dataset sufficiently represents a target population. Additionally, individual DQ measurements of values, registries and attributes can be aggregated to provide summarized DQ results for each axis of the dataset, e.g., the percentage of missing data or inconsistent registries.

Table 7.9 shows a proposal for metrics on each dimension-axis pair. Metrics can be defined to be generic or context-specific. Generic DQ metrics can be measured directly from data without any prior knowledge neither of the domain on which it was acquired nor its purpose. Context-specific metrics can only be measured based on the knowledge associated to the context at which data was acquired or given a specific purpose for data. As an example, the generic (completeness,registry) can be a score of the registry missing data, while a context can define which elements are mandatory and which not in order to provide a weighted indicator of completeness.

7.2.6 Discussion

The purpose of the proposed framework is to serve as a reference for the construction of DQ assessment frameworks, procedures or projects (Figure 7.4), with a core defined by DQ measurements. Thus, the proposed measurements described in Table 7.9 serve as a reference for defining custom metrics, either generic or context-specific. Next, possible assessment methods for the DQ metrics are discussed for different DQ dimensions.

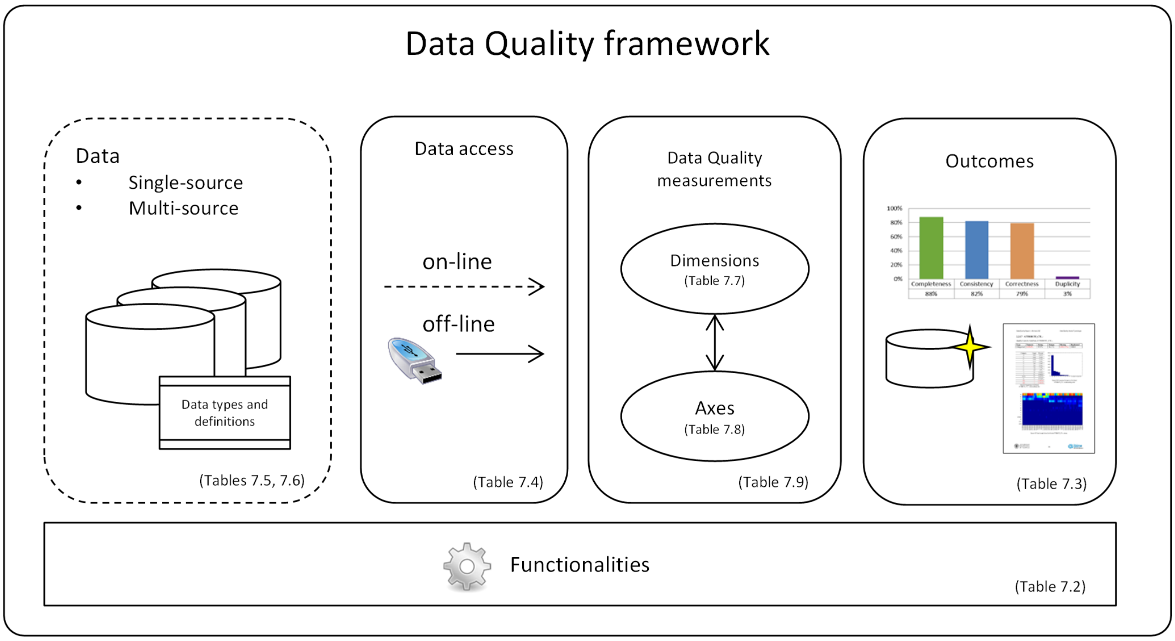


Figure 7.4: Different parts of the proposed DQ framework. Each part is related to the specific table where the content is described.

Previously, the difference between generic and context-specific metric was exemplified with completeness metrics. Other relevant example that shows the versatility of the framework is related to consistency. As defined by the framework, the consistency dimension relies on constraints or rules, which can be domain-independent or domain-specific. First, domain-independent DQ constraints or rules may be defined to apply under any context, e.g., the age of an adult patient must not be negative, or it is impossible for a male to be pregnant. On the other hand, DQ rules may be restricted under specific clinical contexts, e.g., a pediatric department may define an upper age limit above which data is considered inconsistent. Hence, whilst generic rules could be compiled in a real-world knowledge repository to be used in all DQ consistency analysis, context-specific DQ rules may be shared or adapted for the same or similar domains. As defined, consistency can be versatile enough to define DQ rules to be assessed within a single registry or within a population, e.g., to check the simultaneous presence of some conditions within a familiar group.

Table 7.9: Examples of (dimension,axis) measurements. Notes: ‘Degree’ may refer to customized levels of measurement such as accounts, percentages, or specific indicators; ‘NA’ means ‘not applicable’.

	Axis			
Dimension	**Value**	**Registry**	**Attribute**	**Dataset**
Completeness	A value is recorded	Degree of recorded values within a registry or weighted measure by attribute relevance	Degree of recorded values within an attribute	Degree to which the dataset sufficiently represents a target population
Consistency	A value satisfies univariate constraints or rules	Degree to which a registry satisfies multivariate constraints or rules within its values	Degree to which the set of values within an attribute satisfy constraints or rules	Degree to which the set of registries satisfy inter-registry constraints or rules
Uniqueness	NA	Degree of replication of the registry in the dataset	Degree of replicated attributes in the dataset possibly measuring the same information	NA
Correctness	Degree of accuracy and/or precision of a value based on context or other values from the registry, given a gold standard or a probabilistic calculus	Degree of multivariate accuracy and/or precision of a registry based on context or its values, given a gold standard or a probabilistic calculus	Degree of accuracy and/or precision (dispersion, entropy, noise) of the probability distribution of the attribute given a reference	Degree of accuracy and/or precision (dispersion, entropy, noise) of the multivariate probability distribution of the dataset or a set of attributes given a reference
Temporal stability / Timeliness	Degree to which the element/registry remains stable through time on the same patient., or whether an value/registry is up to date		Degree to which the probability distribution of the attribute remains stable through time	Degree to which the multivariate probability distribution of the dataset or a set of attributes remains stable through time
Multi-source stability	Degree to which the value/registry is stable among different sources for the same patient		Degree to which the probability distribution of the attribute is stable among different sources	The degree to which the multivariate probability distribution of the dataset or a set of attributes remains stable among different sources
Contextualization	Degree to which the value/registry is annotated in the context it was acquired		Degree to which the attribute/dataset is contextually and semantically annotated	
Predictive value	Given a decision making purpose, whether the value/registry/attribute/dataset contains the proper information			
Reliability	Degree of reputation of the stakeholders and institutions involved in the acquisition of the value/registry/attribute/dataset			

In the case of uniqueness, we leave to the user the possibility to use any duplicate finding method (Elmagarmid et al, 2007), from simple unique patient ID finding, until record linkage or entity resolution methods, e.g., to find matches of patients with similar demographic data. It can be observed in the table that there is no metric specified for the (uniqueness,dataset) pair. Thus, according to the framework, obtaining the set of replicated registries in a dataset will come as the aggregated measure of (uniqueness,registry).

As previously stated, correctness aims to measure the accuracy and precision of data respect to the real-world concept, and not whether it is valid, which is a matter of consistency. Thus, different assessment methods can also be used to such purpose. For value or registry axes, probabilistic models could provide the likelihood of a value or a set of values to be true given a context or other values in the same registry (Hou and Zhang, 1995; Hipp et al, 2001). In the case of attribute or dataset axes, the likelihood of a population is assessed. In such case, it could be obtained based on a reference gold standard distribution, a probabilistic estimation of the distribution, or quality standards obtained from similar populations.

Regarding to temporal and multi-source stability, in a sense they may be considered similar—except the possible definition of temporal stability about whether data is up to date, see Section 2.1.2. In both cases they can be measured over data snapshots, in the first one based on continuous temporal batches, and in the second one on multi-source subsets of data. In this thesis we have proposed DQ assessment methods for both cases based on information geometry and PDF distances.

Contextualization of data is an important dimension since it contributes understanding the meaning of data, not only to humans, but to computer systems such as automated DQ procedures. If the context at which some data values were acquired is not registered, a physician may miss relevant information for the patient care. Additionally, the development of models or hypotheses may lack of relevant knowledge. Analysing the contextualization of data may consist on checking whether context values associated to patient observations are registered. If data comes from standardized EHRs the contextualization validation process may be even simplified (Maldonado et al, 2012).

The predictive value is always associated to a data purpose. The reuse of clinical routine data for research is currently a common situation when research information repositories are not available. Then, when assessing the quality of a dataset aiming to a specific research purpose, some of these data may present a limited value. Automatically detecting such information may be the purpose of these metrics, e.g., based on data and problem semantics, or measuring the information contained by data with respect to a dependent target variable.

Finally, reliability is a dimension which could be defined or not to be measured from data itself. Generally, the reputation associated to stakeholders comes from external knowledge about them or their past results. However, properly contextualized datasets may contain latent information about such stakeholders for estimating such reputation. E.g., a function composed of DQ metrics obtained from a subset of data related to specific hospitals or physicians may represent a degree of its reputation.

Limitations

We complete the discussion in relation to the limitations of this proposal. The proposed framework intends to assess the quality of biomedical data repositories for data reuse. These types of repositories are generally presented in a format ready for its exploitation, generally in a tabular, or Comma-Separated Values (CSV) format (what includes Excel sheets). Indeed, we may also find repositories of structured clinical documents, however, the tabular format is generally the most seen format for data reuse. The proposed framework better fits tabular input data, but it can be as well generalized, or even extended, to fit the DQ assessment of specific properties of structured documents (as we will see in Section 7.3.3). Nevertheless, methods to flatten structured or relational databases into a purpose-specific table for the DQ analysis might be carried out for their analysis.

Other point to remark is the adequateness or coverage of the proposed characteristics of the framework. The proposed framework establishes a model for DQ assessment which can be followed as-is or could be adapted and extended to specific needs. One may use part of the framework, or use all of its parts. The same would apply to the possible extensions. As an example, we mentioned that we focus on practical dimensions to cover the necessary aspects for data reuse. One may miss specific definitions for any of the proposed dimensions found in the literature, or even miss a complete additional dimension. Hence, the proposed framework could be extended with new definitions of dimensions, being compatible with the defined functionalities, outcomes, variables and axes. Several derived applications of parts of the proposed framework and extensions are described in the next section.

7.3 Derived applications

In this section we describe three DQ assessment applications which are established in the proposed general DQ framework. The first application used the theoretical framework of dimensions and axes for the construction of a data quality assured perinatal data repository (Section 7.3.1). The second application takes de definition of the contextualization dimension and proposes a solution for its assurance on the registry axis (single-case assessment) for the data reuse on CDSSs (Section 7.3.2). The third application (Section 7.3.3) consists in the use of the DQ framework to establish the measurements of DQ dimensions in an on-line service for the evaluation and rating of biomedical data repositories.

7.3.1 Data quality assured perinatal repository

The Baby-friendly Hospital Initiative (BHFI) is an effort by the WHO and UNICEF (2009) to implement practices that protect, promote and support breastfeeding. Having materialized these guidelines, and under a specific research project, the Virgen del Castillo Hospital in Yecla, Spain, decided to use the population data from the HIS for monitoring the perinatal clinical activities matching the evidence of the BHFI.

Taking this opportunity, the Virgen del Castillo Hospital decided to build a computational process to extract the data from the EHRs under a quality control mechanism, towards the construction of a quality-assured perinatal repository for data reuse. This reuse included the aforementioned monitoring and research activities.

Aiming to build a generic solution for the construction of infant feeding repositories from birth until two years, independent of the primary EHRs, and based on a theoretical basis for its quality control, the DQ framework proposed in this chapter was used.

Results

The developed data quality assurance process consists of 13 stages to ensure the harmonization, standardisation, completion, de-duplication, and consistency of the dataset content. The quality of the input and output data at each of these steps is controlled according to eight of the DQ dimensions in Table 7.7: predictive value, correctness, duplication, consistency, completeness, contextualization, temporal-stability and spatial-stability, and measured across the axes in Table 7.8. Consequently, in addition to obtaining a quality assured repository at the end of the process, we obtain its DQ meta-information, which allows monitoring the clinical processes under a TDQM methodology.

The process was applied to obtain a quality assured repository from the original EHRs of the Hospital. The initial dataset consisted of 2,048 registries and 223 attributes with information from the perinatal period. The resultant quality assured repository consisted of 1,925 registries and 73 attributes, discarding those elements that are non-informative for the reuse, with redundant data or with non-recoverable DQ problems (see Table 7.10).

To check the effect of the DQ correction procedures applied at each stage of the process, the DQ was measured at the stage input and output data. Table 7.11 shows a selection of these measurements, where a significant improvement in the DQ measurements can be observed.

Table 7.10: Comparison of the number of elements between initial and quality-assured repository of infant feeding of the Virgen del Castillo Hospital

	Initial dataset	Quality-assured repository
Registries	2,048	1,925
Attributes	223	73
Values (observations)	433,308	107,529

Table 7.11: Selection of data quality measurements when applying the developed data quality assurance process to the infant-feeding dataset of the Virgen del Castillo Hospital

DQ problem	Dimension	Affected before procedure	Affected after procedure
Non-informative attributes	Predictive value	64%	0%
Births with miss-assigned forms	Contextualization	8%	0%
Attributes with changes in support (protocol changes)	Temporal stability	19%	0%
Out of range observations	Consistency	0.003%	0%
Unlikely observations	Correctness	0.001%	0%
Incomplete birth registries	Completeness	6%	0%
Replicated observations	Uniqueness	45%	0%
Observations with variability among their replications	Correctness	1%	0%
Registries with inconsistencies among attributes (multivariate)	Consistency	6%	3%

Discussion

This study emphasized the transparency provided by the DQ assessment in biomedical research repositories and explored the applicability of the DQ framework proposed in this thesis in real scenarios. Besides, this work enabled the construction of the first quality-assured repository for the reuse of information on infant feeding in the perinatal period for monitoring healthcare activities and research purposes.

7.3.2 Contextualization of data for their reuse in CDSSs reuse using an HL7-CDA wrapper

Contextualization is one of the main DQ dimensions to be considered when sharing data among multiple sources, since it helps ensuring common semantics of medical concepts and data understanding. When reusing data for research, such as for data mining and knowledge discovery, the contextualization of data provides researchers a better understanding of the variables and the problem in hand, improving their outcomes. In addition, when predictive or knowledge-base models are used in a CDSS along different locations, it is of upmost importance that the CDSS can understand the semantics of the EHRs to be used as input for the model. Hence, a proper contextualization of data is key for such a purpose.

During the development of this thesis, we participated in a project aiming to develop a knowledge-based personal health system for the empowerment of patients with diabetes mellitus. The knowledge-base model was based on the American Diabetes Association guidelines and it was integrated in a telemedicine system to be used as a remote CDSS from several medical institutions. Hence, the CDSS required patient data related to nutrition habits and physical activities, and vital measurements acquired at different moments. The CDSS was aimed to reuse the original data from the different EHRs in order to calculate the patient recommendations and risk assessments. As a consequence, for an input data of a minimum quality, data should be properly contextualized identifying their semantics and contextual metadata. In addition, the results of the CDSS should also be provided in some standardized and contextualized format to be as well understood by the requesting institutions.

According to the objectives of the project, the DQ assessment was not aimed to measure the data contextualization, but to assure it. Hence, we aimed to offer a first step towards DQ assurance for the reuse of EHRs by CDSSs based on their contextualization. The assurance of other single-case DQ dimensions such as completeness or consistency was out of the scope of the project.

Results

The proposed solution to assure the data contextualization was based on standardized input and output data for the CDSS conforming an Health Level 7 Clinical Document Architecture (HL7-CDA) wrapper. HL7-CDA is a standard for the structure and exchange of clinical documents (Dolin et al, 2006). HL7-CDA is approved by ANSI and is currently one of the most widely accepted clinical documents standard. HL7-CDA contains its own vocabulary which provide a first degree of semantics. However, in most cases it must be completed with clinical terminologies in order to provide the required contextualization for a particular scenario. The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) is currently one of the most extended clinical terminologies worldwide. It contains uniquely coded and, in general, unambiguous clinical concepts from most health care domains. Hence, we combined the use of HL7-CDA with SNOMED-CT to provide data with a complete contextualization. According to the recommendations by HL7, we defined the HL7-CDA restrictions in a HL7-CDA Implementation Guide to formally describe the contents of the proposed HL7-CDA documents.

Besides the required contextualization of data towards its reuse quality, we took the opportunity to make a solution for facilitating its use and generalization on most types of knowledge-based (or rule-based) CDSS. Hence, patient data and rule inference results were mapped respectively to and from the CDSS by means of a binding method based on an XML binding file. This way, we provided a non-invasive solution based just on binding standardized data to the rule facts in the knowledge-base and vice-versa by means of a specific knowledge binding and language files. The proposed binding method permits describing the knowledge-base using human readable terms instead of terminology codes, what facilitates the maintenance of the knowledge. Additionally, being the results of the CDSS an independent standardized clinical document, they can present clinical and legal validity.

Figure 7.5 shows the conceptual schema of the proposed approach. On the left side of the figure we can see the input and output clinical documents wrapping the CDSS. These correspond to HL7-CDA documents. An HL7-CDA implementation guide is provided as the standard template for the output documents as well as a recommended template for the input documents. In the center of the figure we can see the binding layer of the method. Its objectives are 1) reading the input CDA document and transforming it into a set of input facts compatible with the inference engine and 2) obtaining the results of the inference rules and transforming them into the output CDA document using the corresponding language terms for textual values. Finally, the right side of the figure represents the inference engine containing its knowledge-base.

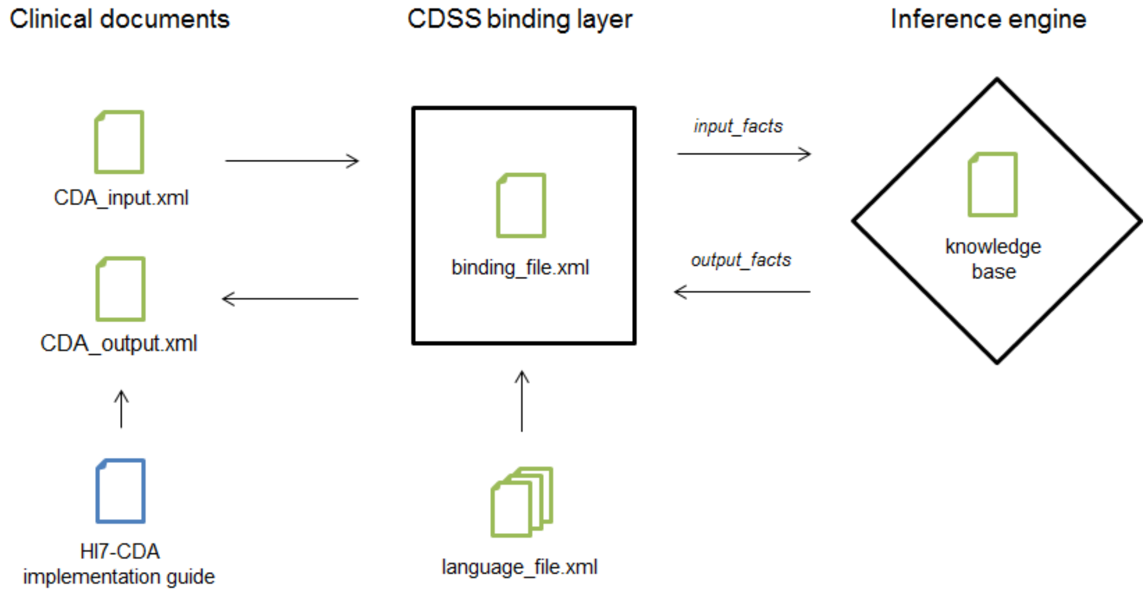
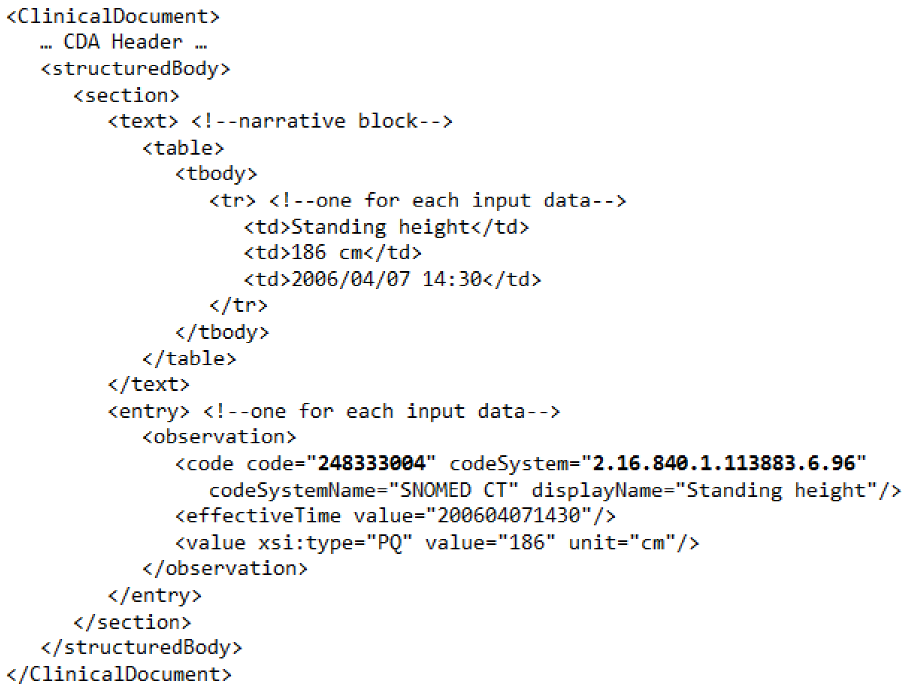


Figure 7.5: Conceptual schema of the proposed solution. Data flow between the three main components and from external inputs is represented by arrows.

Figure 7.6 shows an example of the contextualized input data, including the clinical concept 'Standing height'. First, its semantic is assured being coded with its SNOMED-CT code 2483330004. Second, towards its utilization for the recommendations by the CDSS, it is contextualized with metadata about its acquisition date using the HL7-CDA 'effectiveTime' component. Note that the clinical concept is repeated under 'text' and 'entry' elements. According to the HL7-CDA guidelines, the former represent a narrative block, to be easily translated to human understanding, while the second is aimed to the computer understanding, in our case by the CDSS.

Discussion

It can be to some degree acceptable that the data semantics and contextualization levels may be different across different families of HIS or locations, since they should not need to be interoperable outside the organization. For that reason, we initially focused to the data contextualization. However, the fact of having applied other basic DQ assessment, making data suitable for the data reuse by CDSSs could be an open discussion: should a CDSS always validate the maximum aspects of DQ as possible? should the CDSS assume that such validation was made at original HIS? For reliability reasons, independently of whether DQ is originally assessed, a CDSS would benefit of making its own data quality validation at data input. Basic checks of missing data



```
<ClinicalDocument>
   … CDA Header …
   <structuredBody>
      <section>
         <text> <!--narrative block-->
            <table>
               <tbody>
                  <tr> <!--one for each input data-->
                     <td>Standing height</td>
                     <td>186 cm</td>
                     <td>2006/04/07 14:30</td>
                  </tr>
               </tbody>
            </table>
         </text>
         <entry> <!--one for each input data-->
            <observation>
               <code code="248333004" codeSystem="2.16.840.1.113883.6.96"
                  codeSystemName="SNOMED CT" displayName="Standing height"/>
               <effectiveTime value="200604071430"/>
               <value xsi:type="PQ" value="186" unit="cm"/>
            </observation>
         </entry>
      </section>
   </structuredBody>
</ClinicalDocument>
```

Figure 7.6: Structure and contents of contextualized input data for the CDSS input as a HL7-CDA document. The clinical concept 'Standing height' is here the only input data.

and inconsistencies (be univariate or among various variables) would avoid unreliable CDSS results. In addition, multi-source and temporal variability based checks such as those developed in this thesis may assess whether a CDSS is suitable for the population at which it is intended to be applied, e.g., monitoring the distributions of batches of input data. Further, if the original distribution of the data used to infer the knowledge of a CDSS is available, we could evaluate the outlyingness of single patient cases at input time aiming to provide a degree of possible error of the CDSS result.

7.3.3 Qualize

Concerned with the problematic of data quality, a partnership formed by the UPV and the company VeraTech for Health S.L., started a joint project towards the development of an on-line service for the data quality evaluation and rating of biomedical data repositories, known as Qualize.

Qualize was conceived from the investigation towards the general data quality framework carried out as part of this thesis. Consequently, Qualize relies in the DQ framework of dimensions and axis proposed in this chapter and includes, among others, the methods for multi-source and temporal variability assessment developed in this thesis.

Results

Qualize was designed to provide three main functionalities for the DQ assessment of biomedical data repositories:

1. Evaluate the DQ of repositories based on the DQ dimensions in Table 7.7, providing for each specialized metrics, visualizations, recommendations and generation of DQ reports.

2. Rate the DQ of repositories, positioning them respect to other repositories, with the purpose to encourage the excellence in the field of biomedical data quality.

3. Extract subsets of quality-assured data from repositories, towards the exploitation of sets of valid and reliable data.

Threfore, the core of these functionalities are specific DQ methods associated to the nine DQ dimensions proposed in our framework: multi-source stability, temporal stability, correctness, completeness, uniqueness, consistency, contextualization, predictive value and reliability.

Qualize will count with diverse technologies. First, the multi-source and temporal stability dimensions are being developed based on the analogous variability technologies developed in this thesis. Second, the dimensions of completeness and consistency are being developed based on enriched data archetypes. Third, Qualize is being designed to work with most types of biomedical data, such as plain text or CSV files, health information standards such as ISO EN 13606[f], HL7-CDA Release 2 (Dolin et al, 2006), and openEHR[g]. With respect to the other dimensions, their assessment methods are now under research, e.g., using information geometry approaches derived from the multi-source variability methods.

Other supporting functionalities were also considered. First, the DQ assessment can be fitted-for-purpose according to several functional domains for data including: research, monitoring of indicators, quality of healthcare assistance, or healthcare policies. Second, due to the sensitivity of biomedical data, the data access will be guided by privacy and data protection.

Finally, an special emphasis was put on providing the Qualize service with a user-centered design. To this end, it provides a user-friendly and device-independent GUI, with a clean and responsive design based on latest methodologies and technologies such as Material Design[h] and AngularJS [i]. In addition, the DQ assessment results for multi-souce and temporal variability will provide to the users with navigable versions of the exploratory methods developed in this thesis. Figure 7.7 shows an example of the GUI of the current prototype of Qualize.

Discussion

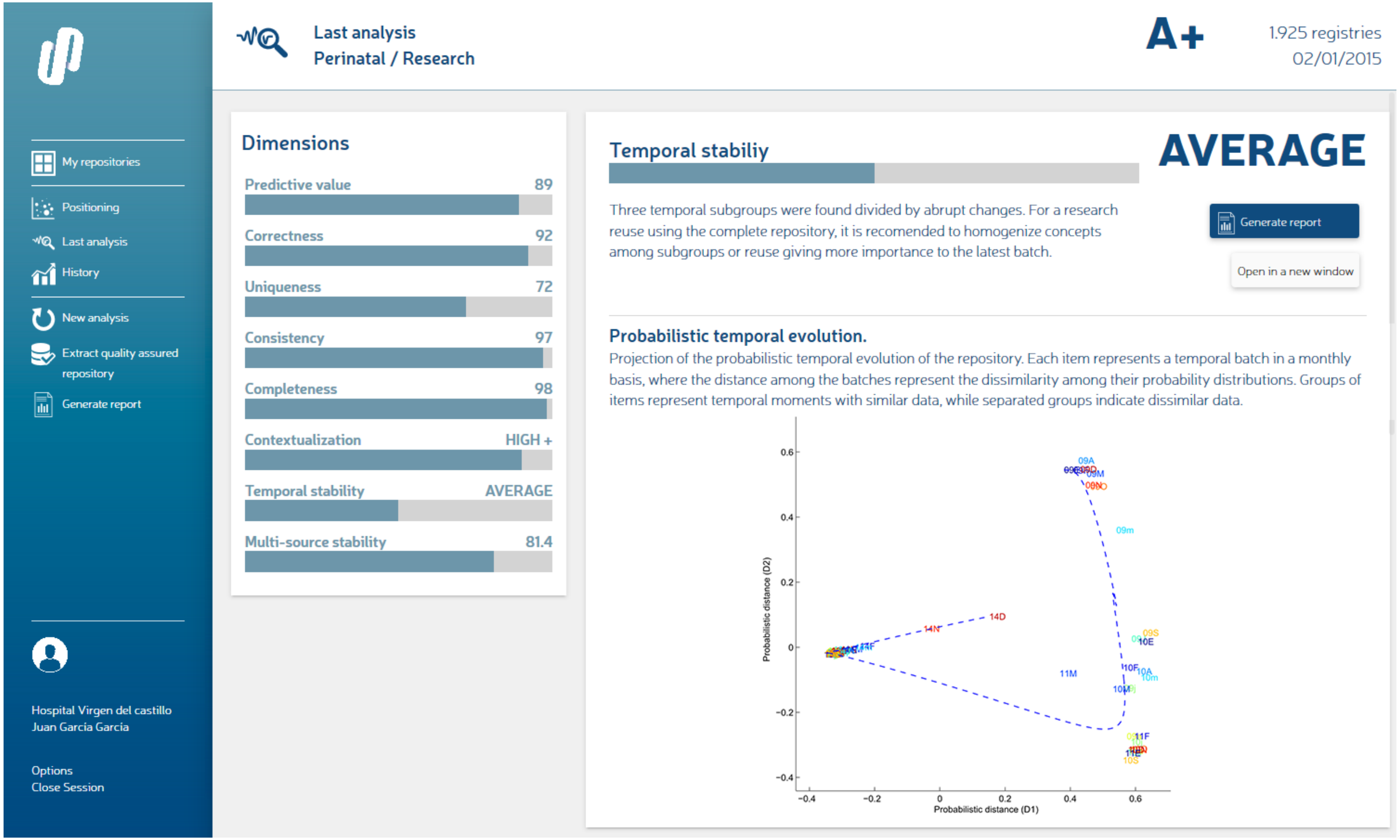
Qualize is an example of how the knowledge and technologies derived from the research carried out in this thesis can be transferred for their exploitation and application to

[f] ISO EN 13606-1:2008 http://www.iso.org/iso/catalogue_detail.htm?csnumber=40784(accessed 2015-09-16)

[g] openEHR Foundation http://openehr.org/ (accessed 2015-09-16)

[h] Google© Material Design http://www.google.com/design/spec/material-design/ (accessed 2015-09-03)

[i] AngularJS by Google© http://angularjs.org/ (accessed 2015-09-03)



Last analysis
Perinatal / Research

A+

1.925 registries
02/01/2015

Hospital Virgen del castillo
Juan Garcia Garcia

Dimensions

Dimension	Value
Predictive value	89
Correctness	92
Uniqueness	72
Consistency	97
Completeness	98
Contextualization	HIGH +
Temporal stability	AVERAGE
Multi-source stability	81.4

Temporal stabiliy

AVERAGE

Three temporal subgroups were found divided by abrupt changes. For a research reuse using the complete repository, it is recomended to homogenize concepts among subgroups or reuse giving more importance to the latest batch.

Probabilistic temporal evolution.

Projection of the probabilistic temporal evolution of the repository. Each item represents a temporal batch in a monthly basis, where the distance among the batches represent the dissimilarity among their probability distributions. Groups of items represent temporal moments with similar data, while separated groups indicate dissimilar data.

Figure 7.7: Example of the GUI of the current prototype of Qualize. The results of the temporal stability DQ dimension assessment of a real perinatal repository are shown.

real problems. With the Qualize service, we intend to provide more value to the healthcare system towards improving the value of their data repositories. Based on the DQ assessment functionalities of the service, we expect to improve the validity and reliability of biomedical data for its reuse in healthcare, strategic, managerial and scientific decision making. The service additionally aims to help discovering which software modules or stages in the clinical workflow are generating DQ problems, reduce the costs of data preparation previous their reuse, certifying the data quality of repositories, and comparing the levels of DQ and the maturity of DQ processes among repositories and institutions.

Chapter 8

Concluding remarks and recommendations

This chapter summarizes the main concluding remarks and recommendations derived from this thesis. This finalizes the work carried out in this thesis, while provides the insights for the continuity of scientific research and development directions based on it.

Part of the provided recommendations were published in the book chapter by Zurriaga et al (2015)—thesis contribution P7.

8.1 Concluding remarks

The existence of large biomedical data repositories with an assured data quality is becoming a reality thanks to the increasing number of open data, data-sharing infrastructures and data quality research. In this thesis we have mainly contributed to the assessment of two data quality problems which are of special importance in multi-source repositories acquired during long periods of time: the variability in data distributions among sources and through time. To this end, we have defined and developed different methods for the assessment of multi-source and temporal variability based on Information Theory and Geometry. The developed methods overcome common problems of classical methods on Big Data sets of multi-modal, multi-type and multi-variate data.

This thesis have contributed to the scientific state-of-the-art in the fields of Medical Informatics, Statistics and Probability, Information Systems, Data Mining and Biomedical Engineering. This is evidenced with the publications derived from this thesis in top-ranked journals and international conferences. In addition, the developed methods have been compiled in a registered software package which facilitates its reuse on further case studies as well as its industrialization.

The specific concluding remarks of this thesis are listed as follows.

CR1 Having reviewed the state-of-the-art in data quality methods (Section 2.1), we found little attention to solutions for assessing and measuring the multi-source and temporal variability. Our own research and various systematic reviews carried out in the literature confirmed that methods to address data quality are heterogeneous. In general, the so-called dimensions provide definitions of what aspects of data quality must be addressed, which vary according to the purpose. Despite the differences in the literature, the underlying concepts of dimensions are likely common. Among them, we found concepts related to the concordance of data semantics with integration purposes, the concordance of data to reference gold-standard data, the degree to which data is up-to-date for current purposes, and the monitoring of indicators in classical quality control approaches. These concepts can be to some degree related to the problem of variability of data distributions among multiple sources or through time. However, the relation to probability distributions is not usually studied, and we have not found any studies that implicitly classify temporal and multi-site variability in probability distributions as DQ dimensions nor those that propose a methodological approach to deal with such variabilities as part of DQ assessment procedures.

We have explicitly classified multi-source and temporal variability in data distributions as data quality problems, and recall the importance of their assessment for a proper data reuse in large-scale multi-source biomedical data repositories.

This concluding remark responds to the research question RQ1, covers the objective O1 and was derived from the works in publications P1, P3, and P4.

CR2 Information-theoretic probability distribution distances are a robust basis for building non-parametric, sample-size and variable-type independent methods for comparing distributions. Classical statistical distribution comparisons are generally problem specific. Classical methods are suited to specific types of (generally univariate) variables, such as numerical normally-distributed data (e.g., ANOVA, or MANOVA, its multivariate alternative), non-parametric data (e.g., Kolmogorov-Smirnov test), or categorical data (e.g., Chi-square test). Besides, not all the measured statistics satisfy the properties of a distance, and the results of statistical tests of hypothesis are generally affected by large sample sizes.

Information-theoretic distances have been the basis for the methods developed in this thesis, which have permitted making multi-source and temporal variability methods suitable to Big Data sets of multi-modal, multi-type and multi-variate data. Specifically, we selected the Jensen-Shannon distance (the Hellinger distance may have been used with similar properties) for satisfying the properties of a distance, being directly computed from the Kullback-Leibler divergence, and for being bounded between zero and one to make the developed methods comparable.

This concluding remark responds to the research question RQ4, covers the objective O2, and was derived from the work in publication P2.

CR3 A method for the assessment of the multi-source variability of biomedical data distributions has been developed providing (1) a metric for measuring the global probabilistic variability among multiple-sources, the GPD, (2) a metric for measuring the outlyingness of a data source with respect to the central tendency, the

SPO, and (3) an exploratory visualization for the inter-source dissimilarity, the multi-source variability simplex. The GPD stands as a metric equivalent to the notion of a probabilistic standard deviation of a set of PDFs to their central tendency. Therefore, as demonstrated in the metric evaluation and the case studies, the GPD can be used as a DQ metric to control the degree of concordance among the distributions of multiple sources. On the other hand, the SPO metric has demonstrated to be a useful indicator to detect data sources with anomalous, biased, or isolated data behaviour. The resultant exploratory visualization has demonstrated in the case studies to be a effective tool to explore the inter-source concordance among distributions, which permits rapidly detecting isolated data behaviours in single data sources or detecting subgroups of data sources with closer distributions.

This concluding remark responds to the research question RQ2, covers the objective O3, and was derived from the work in publication P3.

CR4 A method for the assessment of the temporal variability of biomedical data distributions has been developed providing (1) an visualization plot to explore the temporal evolution and behaviour of distributions, the IGT plot, and (2) a statistical process control algorithm for quantitatively monitoring the variability of data distributions through time, the PDF-SPC. The IGT plot projects in 2D the non-parametric statistical manifold of the set of distributions extracted from dividing the repository into temporal batches, with the advantage of knowing the temporal connection among them. Hence, in the method evaluation and in the case studies, the IGT-plot has demonstrated to capture different types of differences in data distributions through time, namely gradual, abrupt and recurrent changes. On the other hand, the PDF-SPC has shown to be a quantitative complementary method to the IGT plot, which permitted automatically firing warning or out-of-control states according to the degree of temporal variability of the data acquisition process.

This concluding remark responds to the research question RQ3, covers the objective O4, and was derived from the work in publication P4.

CR5 The multi-source and temporal variability methods developed in this thesis, as well as their common probabilistic basis, have been evaluated and validated with simulated benchmarks and real case studies. The selection of proper distances for comparing distributions was evaluated by simulating different types of multi-modal, multi-type and multi-variate distributions and possible changes between them. The multi-source variability method was evaluated during its design using a simulated benchmark for the target features of the method and in a real problem comparing the multiple sources of the UCI Heart Disease dataset. The temporal variability method was evaluated during its design using two variables of the real US NHDS dataset, and simulating temporal changes over them. Finally, both methods were additionally validated on the cases of study described in Chapter 6 including an exhaustive evaluation in the Public Health Mortality or the Region of Valencia, and other evaluations in the Cancer Registry of the Region of Valencia, a National Breast Cancer dataset, and an In-Vitro Fertilization dataset.

These evaluations have demonstrated the usefulness of the developed methods for assessing and controlling the variability of data distributions among sources and through time, first, in conditions at which classical statistical methods are not suitable, second, providing novel quantitative and qualitative information to the date not available with other methods, and, third, resulting in a generic approach suitable for different types of datasets or biomedical data reuse problems.

This concluding remark responds to the research question RQ4, covers the objective O5, and was derived from the work in publications P2, P3, P4 and P5.

CR6 A software containing the methods and algorithms for multi-source and temporal variability developed in this thesis has been developed and registered in the technological offer of the UPV. This includes functions for obtaining the GPD and SPO metrics, multi-source variability simplex visualization, IGT plots, PDF-SPC and temporal heat maps. The functions are built to be generic, that is, to be used on any type of non-parametric distribution, with different types of variables, in a uni- or multivariate setting, and even mixing different types of variables. Additionally, an algorithm for reading Big Data files in streaming and using temporal landmarks to set temporal batches has been included, which incrementally estimates the non-parametric probability distributions to be analysed.

This software, provided as a MATLAB framework, establishes a novel suite of DQ metrics and data profiling tools for the systematic management of multi-source data-sharing infrastructures and multi-source research datasets, as well as for the data understanding and preparation for data mining and knowledge discovery tasks. A systematic approach for assessing the multi-source and temporal variability has been proposed based on the experiences of the use of the methods in the cases of study. Finally, we want to recall that the methods proposed in this thesis could be used as well in traditional data analysis problems, and complemented with other methods such as clustering algorithms to provide more light on the relationships among data sources and time periods in their statistical manifolds.

This concluding remark responds to the research question RQ5, covers the objective O6 and is related to the software contribution S1.

CR7 In addition to the development of the multi-source and temporal variability methods, we have aimed to establish the basis of a general framework for the evaluation of DQ in biomedical data repositories. The objective of this task was to open the path to further research about DQ dimensions and facilitate the industrialization of their assessment methods. As a consequence, in this thesis we have presented two of the outcomes towards such an objective. First, we developed a method to ensure the proper contextualization of biomedical data as the input and output of CDSS based on the standardization of data concepts using the HL7-CDA clinical documents standard. Contextualization is one of the main DQ dimensions to be considered when sharing data among multiple sources, since it helps ensuring a common semantics of medical concepts and data understanding. Second, we have tried to define a common theoretical framework for DQ assessment. It is based on the definition of nine DQ dimensions, aiming to cover the most important dimensions to our opinion from the literature, which can be measured in different axis of the dataset, namely through registries, attributes, single-values, full dataset, multi-source and through time. Several examples for these measurement possibilities were discussed.

This concluding remark responds to research question RQ5, covers the objective O6 and is related to the software contributions S2 and S3.

8.2 Recommendations

The objectives of this thesis were motivated, first, by the background and recommendations given from the years of experience of the IBIME research group, including the author and advisors, in biomedical data analysis. And second, by the global necessity of accessing valid and reliable biomedical data for its reuse in research or decision making, as justified in the scientific state-of-the-art and Big Data tendencies.

As such, continuing with the research cycle, the developed methods and research findings in this thesis can establish the starting point of further research branches based on them, in addition to further technological developments. The following recommendations are suggested.

R1 Even when semantic and integration aspects are solved in large multi-site data sharing infrastructures, probabilistic variability may still be present in data, which may entail different data reuse problems. Unmanaged multi-site and temporal variability may lead to inaccurate or unreproducible research results, or suboptimal decisions. We suggest incorporating the assessment of data temporal and multi-site probabilistic variability in systematic DQ procedures. In the case of multi-site repositories we advocate for assessing their ‘probabilistic interoperability’ based on the GPD and SPO metrics, and the multi-source variability simplex visualization.

R2 The GPD and SPO metrics can be used to provide a quantitative assessment of the multi-source variability of biomedical data repositories, i.e., as data quality metrics. The construction of a metric for the temporal variability dimension can be studied as well based on the presented temporal variability methods. Possible approaches may be either defining heuristics on the changes found in the statistical manifold which originates the IGT plot, e.g., based on the number of temporal subgroups, or based on the changes of state detected by the PDF-SPC algorithm. Besides, it is important to study to which degree changes are normal or expected, such as the gradual changes due to normal environmental changes.

R3 In addition to act as DQ assessment methods, the GPD and SPO metrics can be used as statistical methods for the assessment of differences among samples, with the advantage of being non-parametric, multi-variate, and variable type and sample size independent. Concretely, the GPD can be interpreted (similarly to the Jensen-Shannon distance) as the degree of the percentage of overlapping among the analysed PDFs, i.e., $GPD = 0$ indicates exact distributions, while $GPD = 1$ indicates completely disjoint (or separable) distributions. Therefore, towards facilitating the interpretation of the GPD and SPO to further purposes, a study for their characterization on several problems, relying on many examples and comparing them with several classical statistical tests, should be carried out. As an example, we are currently using the GPD as a metric to evaluate the separability among the tissues of automatically segmented brain tumours based on the distributions of different quantitative MRI attributes of the tissues.

R4 Other possible use for the multi-source variability method to be investigated is for feature selection in classification and regression. In the case of classification, the different classes to be predicted in the dependent variable could act as the source for the GPD. Hence, the independent variables can be divided in sub-distributions based on the class, and compute the GPD metric with them. Hence, the higher the GPD the more separable the classes are with respect to the measured variable. This can as well be extended to the multivariate case, what may be used as the basis for the predictive value DQ dimension as defined in Chapter 7.

R5 The proposed method for temporal variability assessment based on Information Geometry, the IGT-plot, has contributed to the state-of-the-art of change detection and characterization, opening many possibilities for further research. The first next step would be investigating the use of Functional Data Analysis (Ramsay and Silverman, 2005) to model the probabilistic temporal evolution of distributions through the statistical manifold. This could facilitate the characterization of changes based on prototype curves, and to predict the future state of parametric and non-parametric distributions based on a single curve parameter (e.g., to predict the future state of a mixture of distributions with unfixed number of components and parameters).

R6 The multi-source and temporal variability methods, specially the temporal variability one, are based on the non-parametric Information Geometry of data distributions. Other possibilities of Information Geometry remain to be studied, such as using exponential families to rely on a generic parametric model, what may avoid the requirement of using non-parametric embedding techniques, or investigating the capabilities of other unbounded PDF distances. The Information Geometry field is a recognised hard field of study, however, a deeper understanding of it may open many further possibilities.

R7 One common limitation of many data mining, statistical and knowledge discovery studies is how the dimensionality of data complicates the modelling capabilities, widely known as the curse of dimensionality. In a similar way, an analogous problem to dimensionality is the main limitation found in the methods developed in this thesis. This is related to the number of bins at which non-parametric distribution histograms are estimated, mainly affecting the estimation of categorical data, but also when discretizing continuous or mixing types of variables. Hence, the larger the number of bins, the wider the probabilistic space at which individuals can be positioned. This may specially affect when counting with small sample sizes. Hence, it is expected a higher default noise (or a maximum entropy) in models with a larger number of bins. As a consequence, measuring bin-based non-parametric distances among these distributions generally leads to default larger distances. The main drawback for this problem to the methods is that the comparability among metrics and results is to some degree reduced when the number of bins is different. Hence, improving the comparability capabilities when using different number of bins is an important work to be studied.

R8 Other problem related to the analysis of multivariate data is that, when comparing distributions, large differences in specific variables may marginalize smaller but possibly important differences in multivariate interactions. Hence, as in an univariate comparisons large differences in individual variables will be likely found, when these variables have a real, although minor, interaction with other variable, the former univariate change would likely get all the weight in the comparison. As a consequence, more focus should be put in multivariate interactions to remove individual variable effects, e.g., using mutual information.

R9 In the technological aspect, the methods developed in this thesis could be integrated into a graphical user interface which facilitates their systematic use. Concretely, users could obtain the metrics and visualizations dynamically navigating through variables, data sources, and temporal periods of their datasets, following the systematic approach proposed in Section 7.1.1. Additionally, the system could be connected to a database for the automatic monitoring of variability, and the developed automatic reporting methods could be integrated into such software. Further, in some situations Big Data sampling methods could be used to optimize the efficiency of the analyses.

R10 The general framework for DQ assessment proposed in Chapter 7 opens the possibility to define new DQ metrics and methods for the proposed dimensions and axes. Such framework is currently being utilized as the base for an industrial development aimed to the DQ evaluation and DQ rating system, being developed in a joint action by the IBIME research group and the technological company VeraTech for Health S.L.

R11 To ensure the highest levels of DQ and continuously improve data management procedures (e.g., data acquisition or processing), organisational DQ assurance protocols should be established by those organizations which store, process or use biomedical data. DQ assurance protocols combine activities at different levels, from the design of the information system, the user training in DQ, to a continuous DQ control and data curation. These DQ activities can be managed by means of standardized methodologies, for example, based on the Total Quality Management process improvement methodology (Wang, 1998; Röthlin, 2010; Sebastian-Coleman, 2013) or on the ISO-8000 standard.[j] The methods for multi-source and temporal variability assessment developed in this thesis may be used as part of those protocols for the continuous DQ control and multi-site audit which, as demonstrated in this thesis, would be able to provide information about the several DQ aspects reflected in data distributions. As part of a cyclic methodology, the outcomes provided by the DQ control may allow defining strategies to prevent and correct DQ problems from their acquisition, for example when manually registering patient observations, until their reuse for research or population studies. Finally, we remark that given the trends in large-scale data-sharing projects, leading to open, Big Data repositories of biomedical data, the importance of DQ assessment and assurance procedures will become even higher, representing a success factor.

[j] ISO/TS 8000-1:2011 Data Quality - Part 1: Overview http://www.iso.org/iso/catalogue_detail.htm?csnumber=50798 (accessed 2015-09-17)

Bibliography

Aggarwal C (2003) A framework for diagnosing changes in evolving data streams. In: ACM SIGMOD Conference, pp 575–586

Ali SM, Silvey SD (1966a) A General Class of Coefficients of Divergence of One Distribution from Another. Journal of the Royal Statistical Society Series B (Methodological) 28(1):131–142

Ali SM, Silvey SD (1966b) A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society Series B (Methodological) pp 131–142

Alpar P, Winkelsträter S (2014) Assessment of data quality in accounting data with association rules. Expert Syst Appl 41(5):2259–2268

Amari SI (2001) Information geometry on hierarchy of probability distributions. Information Theory, IEEE Transactions on 47(5):1701–1711

Amari SI, Nagaoka H (2007) Methods of Information Geometry (Translations of Mathematical Monographs). American Mathematical Society

Applegate D, Dasu T, Krishnan S, Urbanek S (2011) Unsupervised clustering of multidimensional distributions using earth mover distance. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, USA, KDD '11, pp 636–644

Arias E (2014) United states life tables, 2009. National Vital Statistics Reports 62(7)

Arts DGT, de Keizer NF, Scheffer GJ (2002) Defining and improving data quality in medical registries: A literature review, case study, and generic framework. Journal of the American Medical Informatics Association 9(6):600–611

Aspden P, Corrigan JM, Wolcott J, Erickson SM (2004) Patient safety: achieving a new standard for care

Asuncion A, Newman D (2007) UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences http://archive.ics.uci.edu/ml/ (Accessed 27/03/2014)

Basseville M (2010) Divergence measures for statistical data processing

Basseville M, Nikiforov IV (1993) Detection of Abrupt Changes: Theory and Application. Prentice-Hall, Inc., Upper Saddle River, NJ, USA

Batini C, Cappiello C, Francalanci C, Maurino A (2009) Methodologies for data quality assessment and improvement. ACM Computing Surveys 41(3):1–52

Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(8):1798–1828

Berger JO (2003) Could fisher, jeffreys and neyman have agreed on testing? Statistical Science 18(1):1–32

Bernet L, García-Gómez JM, Cano Muñoz R, Piñero A, Ramírez AK, Rodrigo M, de la Cámara de las Heras JM, Burgués O, Ruiz I, Tormos B (2015) Modelo predictivo multiparamétrico del estatus axilar en pacientes con cáncer de mama: carga tumoral total y perfil molecular. estudio multicéntrico. Revista de Senología y Patología Mamaria 28(3):96–104

Biau DJ, Kernéis S, Porcher R (2008) Statistics in brief: The importance of sample size in the planning and interpretation of medical research. Clinical Orthopaedics and Related Research 466(9):2282–2288

Borg I, Groenen PJF (2010) Modern Multidimensional Scaling: Theory and Applications. Springer

Bowman AW, Azzalini A (1997) Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations (Oxford Statistical Science Series). Oxford University Press, USA

Brandes U, Pich C (2007) Eigensolver methods for progressive multidimensional scaling of large data. In: Kaufmann M, Wagner D (eds) Graph Drawing, Springer Berlin Heidelberg, Lecture Notes in Computer Science, vol 4372, pp 42–53

Bray F, Parkin DM (2009) Evaluation of data quality in the cancer registry: Principles and methods. part i: Comparability, validity and timeliness. European Journal of Cancer 45(5):747–755

Brazdil PB (ed) (2009) Metalearning: applications to data mining. Cognitive technologies, Springer, Berlin

Bresó A, Sáez C, Vicente J, Larrinaga F, Robles M, García-Gómez JM (2015) Knowledge-based personal health system to empower outpatients of diabetes mellitus by means of p4 medicine. In: Fernández-Llatas C, García-Gómez JM (eds) Data Mining in Clinical Medicine, vol 1246, Springer New York, pp 237–257

Brockwell P, Davis R (2009) Time Series: Theory and Methods. Springer Series in Statistics, Springer

Budka M, Gabrys B, Musial K (2011) On accuracy of pdf divergence estimators and their applicability to representative data sampling. Entropy 13(7):1229–1266

Cali A, Calvanese D, Giacomo GD, Lenzerini M (2004) Data integration under integrity constraints. Information Systems 29(2):147 – 163, the 14th International Conference on Advanced Information Systems Engineering (CAiSE'02)

Carter KM, Raich R, Finn WG, Hero AO (2008) Fine: Fisher information non-parametric embedding. arXiv preprint arXiv:08022050

Carvalho AR, Tavares JMRS, Principe JC (2013) A novel nonparametric distance estimator for densities with error bounds. Entropy 15(5):1609–1623

Cayton L (2005) Algorithms for manifold learning. Univ of California at San Diego Tech Rep

Cesario SK (2002) The "Christmas Effect" and other biometeorologic influences on childbearing and the health of women. J Obstet Gynecol Neonatal Nurs 31(5):526–535

Chakrabarti K, Garofalakis M, Rastogi R, Shim K (2001) Approximate query processing using wavelets. The VLDB Journal 10(2-3):199–223

Chen H, Hailey D, Wang N, Yu P (2014) A review of data quality assessment methods for public health information systems. International Journal of Environmental Research and Public Health 11(5):5170–5207

Choi OH, Lim JE, Na HS, Baik DK (2008) An efficient method of data quality using quality evaluation ontology. In: Convergence and Hybrid Information Technology, 2008. ICCIT '08. Third International Conference on, vol 2, pp 1058–1061

Choquet R, Qouiyd S, Ouagne D, Pasche E, Daniel C, Boussaïd O, Jaulent MC (2010) The information quality triangle: a methodology to assess clinical information quality. Stud Health Technol Inform 160:699–703

Cichocki A, Cruces S, Amari Si (2011) Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. Entropy 13(1):134–170

Conover WJ (1999) Practical nonparametric statistics, 3rd edn. Wiley series in probability and statistics. Applied probability and statistics section, Wiley

Cornuéjols A (2010) On-Line Learning: Where Are We So Far? In: May M, Saitta L (eds) Ubiquitous Knowledge Discovery, Lecture Notes in Computer Science, vol 6202, Springer Berlin Heidelberg, pp 129–147

Costa S, Santos S, Strapasson J (2005) Fisher information matrix and hyperbolic geometry. In: Information Theory Workshop, 2005 IEEE

Cover TM, Thomas JA (1991) Elements of Information Theory, 99th edn. Wiley-Interscience

Cranmer K (2014) Visualizing information geometry with multidimensional scaling. http://nbviewer.ipython.org/github/cranmer/play/blob/master/manifoldLearning/GaussianInformationGeometryEmbedding.ipynb (Accessed 26/09/2015)

Croitoru M, Hu B, Dasmahapatra S, Lewis P, Dupplaw D, Gibb A, Julia-Sape M, Vicente J, Saez C, Garcia-Gomez JM, Roset R, Estanyol F, Rafael X, Mier M (2007) Conceptual graphs based information retrieval in HealthAgents. IEEE, pp 618–623

Cruz-Correia RJ, Pereira Rodrigues P, Freitas A, Canario Almeida F, Chen R, Costa-Pereira A (2010) Data quality and integration issues in electronic health records. In: Information Discovery On Electronic Health Records, V. Hristidis (ed.), pp 55–96

Csiszár I (1967) Information-type measures of difference of probability distributions and indirect observations. Studia Scientiarum Mathematicarum Hungarica 2:299–318

Csiszár I (1967) Information-type measures of difference of probability distributions and indirect observations. Studia Sci Math Hungar 2:299–318

Csiszár I (1972) A class of measures of informativity of observation channels. Periodica Mathematica Hungarica 2(1-4):191–213

Csiszár I, Shields PC (2004) Information theory and statistics: A tutorial. Now Publishers Inc

Dargaud Y, Wolberg AS, Luddington R, Regnault V, Spronk H, Baglin T, Lecompte T, Cate HT, Negrier C (2012) Evaluation of a standardized protocol for thrombin generation measurement using the calibrated automated thrombogram: An international multicentre study. Thrombosis Research 130(6):929–934

Dasu T, Loh JM (2012) Statistical distortion: consequences of data cleaning. Proc VLDB Endow 5(11):1674–1683

Dasu T, Krishnan S, Lin D, Venkatasubramanian S, Yi K (2009) Change (detection) you can believe in: Finding distributional shifts in data streams. In: Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII, Springer-Verlag, Berlin, Heidelberg, IDA '09, pp 21–34

De Leeuw J (1993) Fitting distances by least squares. Tech Rep No 130, Interdivisional Program in Statistics, UCLA

Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39(1):1–38

Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, Guppy KH, Lee S, Froelicher V (1989) International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology 64(5):304–310

Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, Shabo Shvo A (2006) HL7 Clinical Document Architecture, Release 2. Journal of the American Medical Informatics Association : JAMIA 13(1):30–39

Elmagarmid AK, Ipeirotis PG, Verykios VS (2007) Duplicate record detection: A survey. Knowledge and Data Engineering, IEEE Transactions on 19(1):1–16

Endres D, Schindelin J (2003) A new metric for probability distributions. IEEE Transactions on Information Theory 49(7):1858–1860

English LP (2006) IQ Characteristics: Information Definition Quality. The Information and Data Quality Newsletter 2(2)

Etcheverry L, Marotta A, Ruggia R (2010) Data quality metrics for genome wide association studies. In: Database and Expert Systems Applications (DEXA), 2010 Workshop on, pp 105–109

Fisher RA (1974) The design of experiments., 9th edn. Hafner Press

Fuster-Garcia E, Navarro C, Vicente J, Tortajada S, García-Gómez JM, Sáez C, Calvar J, Griffiths J, Julià-Sapé M, Howe Fa, Pujol J, Peet AC, Heerschap A, Moreno-Torres A, Martínez-Bisbal MC, Martínez-Granados B, Wesseling P, Semmler W, Capellades J, Majós C, Alberich-Bayarri A, Capdevila A, Monleón D, Martí-Bonmatí L, Arús C, Celda B, Robles M (2011) Compatibility between 3t 1h SV-MRS data and automatic brain tumour diagnosis support systems based on databases of 1.5t 1h SV-MRS spectra. Magma (New York, NY) 24(1):35–42

Galea S, Ahern J, Karpati A (2005) A model of underlying socioeconomic vulnerability in human populations: evidence from variability in population health and implications for public health. Social Science & Medicine 60(11):2417–2430

Gama J (2010) Knowledge Discovery from Data Streams, 1st edn. Chapman & Hall/CRC

Gama J, Gaber MM (2007) Learning from Data Streams: Processing Techniques in Sensor Networks. Springer

Gama J, Medas P, Castillo G, Rodrigues P (2004) Learning with drift detection. In: Bazzan A, Labidi S (eds) Advances in Artificial Intelligence – SBIA 2004, Lecture Notes in Computer Science, vol 3171, Springer Berlin Heidelberg, pp 286–295

García-Gómez JM, Luts J, Julià-Sapé M, Krooshof P, Tortajada S, Robledo JV, Melssen W, Fuster-García E, Olier I, Postma G, Monleón D, Moreno-Torres n, Pujol J, Candiota AP, Martínez-Bisbal MC, Suykens J, Buydens L, Celda B, Van Huffel S, Arús C, Robles M (2009) Multiproject–multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy. Magnetic Resonance Materials in Physics, Biology and Medicine 22(1):5–18

García–Gómez JM, Tortajada S, Vicente J, Sáez C, Castells X, Luts J, Julià–Sapé M, Juan–Císcar A, Van Huffel S, Barceló A, Ariño J, Arús C, Robles M (2007) Genomics and metabolomics research for brain tumour diagnosis based on machine learning. In: Sandoval F, Prieto A, Cabestany J, Graña M (eds) Computational and Ambient Intelligence, vol 4507, Springer Berlin Heidelberg, pp 1012–1019

Gehrke J, Korn F, Srivastava D (2001) On computing correlated aggregates over continual data streams. SIGMOD Rec 30(2):13–24

Goodman SN (1999) Toward evidence-based medical statistics. 1: The p value fallacy. Annals of Internal Medicine 130(12):995

Greenland S (2011) Null misinterpretation in statistical testing and its impact on health risk assessment. Preventive Medicine 53(4):225–228

Greenland S, Poole C (2013) Living with statistics in observational research. Epidemiology 24(1):73–78

Guha S, Shim K, Woo J (2004) Rehist: Relative error histogram construction algorithms. In: Proceedings of the 30th VLDB conference, pp 300–311

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. The Journal of Machine Learning Research 3:1157–1182

Hahn GJ, Shapiro SS (1968) Statistical models in engineering. In: Statistical models in engineering, John Wiley & Sons

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: An update. SIGKDD Explor Newsl 11(1):10–18

Halsey LG, Curran-Everett D, Vowler SL, Drummond GB (2015) The fickle p value generates irreproducible results. Nature methods 12(3):179–185

Han B, Comaniciu D, Zhu Y, Davis L (2004) Incremental density approximation and kernel-based bayesian filtering for object tracking. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol 1, pp I–638–I–644 Vol.1

Han J, Kamber M, Pei J (2012) Data mining: concepts and techniques. Elsevier : Morgan Kaufmann

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. Springer Series in Statistics, Springer New York, New York, NY, USA

Hazewinkel M (1988) Encyclopaedia of mathematics: an updated and annotated translation of the Soviet Mathematical encyclopaedia

Heinrich B, Kaiser M, Klier M (2007) How to measure data quality - a metric based approach. In: In appraisal for: International Conference on Information Systems

Heinrich B, Klier M, Kaiser M (2009) A procedure to develop metrics for currency and its application in CRM. Journal of Data and Information Quality 1(1):1–28

Hero AO, Ma B, Michel O, Gorman J (2001) Alpha-divergence for classification, indexing and retrieval. Communication and Signal Processing Laboratory, Technical Report CSPL-328, U Mich

Hershey J, Olsen P (2007) Approximating the kullback leibler divergence between gaussian mixture models. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, vol 4, pp 317–320

Hipp J, Güntzer U, Grimmer U (2001) Data quality mining-making a virute of necessity. In: DMKD

Hofmann M, Klinkenberg R (2013) RapidMiner: Data Mining Use Cases and Business Analytics Applications. Chapman & Hall/CRC

Hou WC, Zhang Z (1995) Enhancing database correctness: a statistical approach. In: ACM SIGMOD Record, ACM, vol 24, pp 223–232

Howden LM, Meyer JA (2011) Age and Sex Composition: 2010. 2010 Census Briefs US Department of Commerce, Economics and Statistics Administration, US Census Bureau

Hrovat G, Stiglic G, Kokol P, Ojstersek M (2014) Contrasting temporal trend discovery for large healthcare databases. Computer Methods and Programs in Biomedicine 113(1):251–257

Hu B, Croitoru M, Roset R, Dupplaw D, Lurgi M, Dasmahapatra S, Lewis P, Martínez-Miranda J, Sáez C (2011) The HealthAgents ontology: knowledge representation in a distributed decision support system for brain tumours. The Knowledge Engineering Review 26(3):303–328

Ihler, A and Mandel, M (2003) Kernel Density Estimation Toolbox for MATLAB . http://www.ics.uci.edu/~ihler/code/kde.html (Accessed 26/09/2015)

Jantke P (1993) Types of incremental learning. In: AAAI Symposium on Training Issues in Incremental Learning, pp 23–25

Jarman B, Gault S, Alves B, Hider A, Dolan S, Cook A, Hurwitz B, Iezzoni LI (1999) Explaining differences in english hospital death rates using routinely collected data. BMJ 318(7197):1515–1520

Jayram TS (2009) Hellinger strikes back: A note on the multi-party information complexity of and. In: Proceedings of the 12th International Workshop and 13th International Workshop on Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, Springer-Verlag, Berlin, Heidelberg, APPROX '09 / RANDOM '09, pp 562–573

Jeffreys H (1973) Scientific inference. Cambridge University Press

Jeusfeld M, Quix C, Jarke M (1998) Design and analysis of quality information for data warehouses. In: Ling TW, Ram S, Li Lee M (eds) Conceptual Modeling – ER '98, Lecture Notes in Computer Science, vol 1507, Springer Berlin Heidelberg, pp 349–362

Judah, Saul and Friedman, Ted (2014) Gartner Magic Quadrant for Data Quality Tools. Gartner http://www.gartner.com/ (G00261683)

Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF (2012) A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research:. Medical Care 50:S21–S29

Karr AF, Sanil AP, Banks DL (2006) Data quality: A statistical perspective. Statistical Methodology 3(2):137 – 173

Keim DA (2000) Designing pixel-oriented visualization techniques: Theory and applications. IEEE Transactions on Visualization and Computer Graphics 6(1):59–78

Kifer D, Ben-David S, Gehrke J (2004) Detecting change in data streams. In: Proceedings of the Thirtieth international conference on Very large data bases - Volume 30, VLDB Endowment, VLDB '04, pp 180–191

Kim J, Scott CD (2012) Robust kernel density estimation. The Journal of Machine Learning Research 13(1):2529–2565

Klinkenberg R, Renz I (1998) Adaptive information filtering: Learning in the presence of concept drifts. In: Workshop Notes of the ICML/AAAI-98 Workshop Learning for Text Categorization, AAAI Press, pp 33–40

Knatterud GL (2002) Management and conduct of randomized controlled trials. Epidemiologic reviews 24(1):12–25

Knatterud GL, Rockhold FW, George SL, Barton FB, Davis CE, Fairweather WR, Honohan T, Mowery R, O'Neill R (1998) Guidelines for quality assurance in multicenter trials: a position paper. Controlled clinical trials 19(5):477–493

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biological Cybernetics 43(1):59–69

Kruskal J (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29(1):1–27

Kuhn M, Johnson K (2013) Applied predictive modeling. Springer

Kullback S, Leibler RA (1951) On information and sufficiency. The Annals of Mathematical Statistics 22(1):79–86

Lee JA, Verleysen M (2007) Nonlinear dimensionality reduction. Springer

Lee YW, Strong DM, Kahn BK, Wang RY (2002) Aimq: A methodology for information quality assessment. Inf Manage 40(2):133–146

García de León Chocano R, Sáez C, Muñoz-Soler V, García de León González R, García-Gómez JM (2015) Construction of quality-assured infant feeding process of care data repositories: definition and design (Part 1). Computers in Biology and Medicine 67:95–103

García de León Chocano R, Muñoz-Soler V, Sáez C, García de León González R, García-Gómez JM (2016) Construction of quality-assured infant feeding process of care data repositories: construction of the perinatal repository (Part 2). Computers in Biology and Medicine [Accepted]

Liaw S, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, Jalaludin B, Yeo A, Talaei-Khoei A (2013) Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. International Journal of Medical Informatics 82(1):10–24

Liaw ST, Taggart J, Dennis S, Yeo A (2011) Data quality and fitness for purpose of routinely collected data–a general practice case study from an electronic practice-based research network (ePBRN). AMIA Annu Symp Proc 2011:785–794

Liese F, Vajda I (2006) On divergences and informations in statistics and information theory. IEEE Transactions on Information Theory 52(10):4394–4412

Lin J (1991) Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory 37:145–151

Lin M, Lucas HC, Shmueli G (2013) Too big to fail: Large samples and the p -value problem. Information Systems Research 24(4):906–917

Liu H, Song D, Rüger S, Hu R, Uren V (2008) Comparing dissimilarity measures for content-based image retrieval. In: Proceedings of the 4th Asia information retrieval conference on Information retrieval technology, Springer-Verlag, Berlin, Heidelberg, AIRS'08, pp 44–50

Lluch-Ariet M, Estanyol F, Mier M, Delgado C, González-Vélez H, Dalmas T, Robles M, Sáez C, Vicente J, Van Huffel S, Luts J, Arús C, Silveira APC, Julià-Sapé M, Peet A, Gibb A, Sun Y, Celda B, Bisbal MCM, Valsecchi G, Dupplaw D, Hu B, Lewis P (2007) On the implementation of HealthAgents: Agent-based brain tumour diagnosis. In: Annicchiarico R, Cortés U, Urdiales C (eds) Agent Technology and e-Health, Birkhäuser Basel, pp 5–24

Lopes R, Reid I, Hobson P (2007) The two-dimensional kolmogorov-smirnov test. In: XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research, Nikhef, Amsterdam, the Netherlands, April 23-27, 2007

Madnick SE, Wang RY (1992) Introduction to total data quality management (TDQM) research program

Madnick SE, Wang RY, Lee YW, Zhu H (2009) Overview and framework for data and information quality research. Journal of Data and Information Quality 1(1):1–22

Maldonado JA, Costa CM, Moner D, Menárguez-Tortosa M, Boscá D, Giménez JAM, Fernández-Breis JT, Robles M (2012) Using the researchehr platform to facilitate the practical application of the ehr standards. Journal of biomedical informatics 45(4):746–762

Markus HS, Ackerstaff R, Babikian V, Bladin C, Droste D, Grosset D, Levi C, Russell D, Siebler M, Tegeler C (1997) Intercenter agreement in reading doppler embolic signals: A multicenter international study. Stroke 28(7):1307–1310

Mattsson N, Zetterberg H, et al (2010) Lessons from multicenter studies on csf biomarkers for alzheimer's disease. International journal of Alzheimer's disease

McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, Bickel J, Wattanasin N, Gilbert C, Trevvett P, Churchill S, Kohane IS (2013) Shrine: Enabling nationally scalable multi-site disease studies. PLoS ONE 8(3):e55,811

Mitchell TM, Caruana R, Freitag D, McDermott J, Zabowski D (1994) Experience with a learning personal assistant. Commun ACM 37(7):80–91

Müller H, Naumann F, christoph Freytag J (2003) Data quality in genome databases. In: International Conference on Information Quality, pp 269–284

Morimoto T (1963) Markov processes and the h-theorem. Journal of the Physical Society of Japan 18(3):328–331

Mouss H, Mouss D, Mouss N, Sefouhi L (2004) Test of page-hinckley, an approach for fault detection in an agro-alimentary production system. In: Control Conference, 2004. 5th Asian, vol 2, pp 815–818 Vol.2

Murray RL (1999) Basic QC practices. training in statistical quality control for healthcare laboratories. james o. westgard, elsa quam, and trish barry. madison, WI: Westgard quality corporation, 1998, 238 pp., $50.00. ISBN 1-886958-09-2. Clinical Chemistry 45(6):912–912

Nagpaul P (1999) Guide to advanced data analysis using IDAMS software. UNESCO, New Delhi, India

Nash J (1956) The imbedding problem for riemannian manifolds. Annals of Mathematics 63(1):20–63

National Research Council (2011) Explaining Different Levels of Longevity in High-Income Countries. The National Academies Press, Washington, D.C.

Navarro C, Molina JA, Barrios E, Izarzugaza I, Loria D, Cueva P, Sánchez MJ, Chirlaque MD, Fernández L (2013) Evaluación externa de registros de cáncer de base poblacional: la guía REDEPICAN para américa latina. Rev Panam Salud Publica 34(5):337

Ney H, Essen U, Kneser R (1994) On structuring probabilistic dependences in stochastic language modelling. Computer Speech and Language 8(1)

NHDS (2010) United states department of health and human services. centers for disease control and prevention. national center for health statistics. national hospital discharge survey 2008 codebook

NHDS (2014) National Center for Health Statistics, National Hospital Discharge Survey (NHDS) data, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland, available at: http://www.cdc.gov/nchs/nhds.htm

Nielsen F, Garcia V (2009) Statistical exponential families: A digest with flash cards. arXiv preprint arXiv:09114863

Nuzzo R (2014) Statistical errors. Nature 506(13):150–152

Oliveira P, Rodrigues F, Henriques P (2005) A Formal Definition of Data Quality Problems. In: Proceedings of the 2005 International Conference on Information Quality (MIT IQ Conference), Cambridge, MA, USA, MIT

Pagani E, Hirsch JG, Pouwels PJ, Horsfield MA, Perego E, Gass A, Roosendaal SD, Barkhof F, Agosta F, Rovaris M, Caputo D, Giorgio A, Palace J, Marino S, De Stefano N, Ropele S, Fazekas F, Filippi M (2010) Intercenter differences in diffusion tensor MRI acquisition. Journal of Magnetic Resonance Imaging 31(6):1458–1468

Papadimitriou S, Sun J, Faloutsos C (2005) Streaming pattern discovery in multiple time-series. In: Proceedings of the 31st International Conference on Very Large Data Bases, VLDB Endowment, VLDB '05, pp 697–708

Parks HR, Wills DC (2002) An elementary calculation of the dihedral angle of the regular n-simplex. The American Mathematical Monthly 109(8):756–758

Parzen E (1962) On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33(3):1065–1076

Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6 2(11):559–572

Petersen P (2006) Riemannian geometry, 2nd edn. No. 171 in Graduate texts in mathematics, Springer

Pierce EM (2004) Assessing data quality with control matrices. Commun ACM 47(2):82–86

Pipino LL, Lee YW, Wang RY (2002) Data quality assessment. Communications of the ACM 45(4):211–218

Ramsay JO, Silverman BW (2005) Functional data analysis. Springer, New York

Rodrigues PP, Correia RC (2013) Streaming virtual patient records. In: G. Krempl, I. Zliobaite, Y. Wang, G. Forman (Eds.), Real-World Challenges for Data Stream Mining, Otto-von-Guericke, University Magdeburg, pp 34–37

Rodrigues PP, Gama J (2010) A simple dense pixel visualization for mobile sensor data mining. In: Proceedings of the Second International Conference on Knowledge Discovery from Sensor Data, Springer-Verlag, Berlin, Heidelberg, Sensor-KDD'08, pp 175–189

Rodrigues PP, Gama J, Pedroso J (2008) Hierarchical clustering of time-series data streams. Knowledge and Data Engineering, IEEE Transactions on 20(5):615–627

Rodrigues PP, Gama J, Sebastião R (2010) Memoryless fading windows in ubiquitous settings. In: In Proceedings of Ubiquitous Data Mining (UDM) Workshop in conjunction with the 19th European Conference on Artificial Intelligence - ECAI 2010, pp 27–32

Rodrigues PP, Sebastião R, Santos CC (2011) Improving cardiotocography monitoring: a memory-less stream learning approach. In: Proceedings of the Learning from Medical Data Streams Workshop. Bled, Slovenia

Röthlin M (2010) Management of data quality in enterprise resource planning systems, 1st edn. No. Bd. 68 in Reihe: Wirtschaftsinformatik, Eul

Rubner Y, Tomasi C, Guibas L (2000) The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision 40(2):99–121

Sáez C, García-Gómez J, Vicente J, Tortajada S, Esparza M, Navarro A, Fuster E, Robles M, Martí-Bonmatí L, Arús C (2008) A generic decision support system featuring an assembled view of predictive models for magnetic resonance and clinical data. vol 21, p 483

Sáez C, García-Gómez J, Vicente J, Tortajada S, Fuster E, Esparza M, Navarro A, Robles M (2009) Curiam BT 1.0, decision support system for brain tumour diagnosis. vol 22, p 538

Sáez C, García-Gómez J, Vicente J, Tortajada S, Luts J, Dupplaw D, Huffel SV, Robles M (2011) A generic and extensible automatic classification framework applied to brain tumour diagnosis in HealthAgents. The Knowledge Engineering Review 26(3):283–301

Sáez C, García-Gómez JM, Alberich-Bayarri Á, Edo MA, Vaño M, Català-Gregori A, Poyatos C, Mollá E, Martí-Bonmatí L, Robles M (2012a) Clinical validation of the added value of a clinical decision support system for brain tumour diagnosis based on SV 1h MRS: randomized controlled trial of effectiveness and qualitative evaluation

Sáez C, Martínez-Miranda J, Robles M, García-Gómez JM (2012b) Organizing data quality assessment of shifting biomedical data. Stud Health Technol Inform 180:721–725

Sáez C, Bresó A, Vicente J, Robles M, García-Gómez JM (2013a) An HL7-CDA wrapper for facilitating semantic interoperability to rule-based clinical decision support systems. Computer methods and programs in biomedicine 109(3):239–249

Sáez C, Robles M, García-Gómez JM (2013b) Comparative study of probability distribution distances to define a metric for the stability of multi-source biomedical research data. In: Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 3226–3229

Sáez C, Martí-Bonmatí L, Alberich-Bayarri Á, Robles M, García-Gómez JM (2014a) Randomized pilot study and qualitative evaluation of a clinical decision support system for brain tumour diagnosis based on SV 1h MRS: Evaluation as an additional information procedure for novice radiologists. Computers in Biology and Medicine 45:26–33

Sáez C, Robles M, Garcia-Gomez JM (2014b) Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. Statistical Methods in Medical Research Published Online First [In Press]

Sáez C, Rodrigues PP, Gama J, Robles M, García-Gómez JM (2015) Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. Data Mining and Knowledge Discovery 29(4):950–975

Sáez C, Zurriaga O, Pérez-Panadés J, Melchor I, Robles M, García-Gómez JM (2016) Applying probabilistic temporal and multi-site data quality control methods to a public health mortality registry in Spain: A systematic approach to quality control of repositories. Journal of the American Medical Informatics Association [Accepted]

Sayer DC, Goodridge DM (2007) Pilot study: assessment of interlaboratory variability of sequencing-based typing DNA sequence data quality. Tissue Antigens 69:66–68

Sebastian-Coleman L (2013) Measuring data quality for ongoing improvement: a data quality assessment framework. Morgan Kaufmann

Sebastião R, Gama J (2009) A study on change detection methods. In: 4th Portuguese Conf. on Artificial Intelligence

Sebastião R, Gama J, Rodrigues P, Bernardes J (2010) Monitoring incremental histogram distribution for change detection in data streams. In: Gaber M, Vatsavai R, Omitaomu O, Gama J, Chawla N, Ganguly A (eds) Knowledge Discovery from Sensor Data, Lecture Notes in Computer Science, vol 5840, Springer Berlin Heidelberg, pp 25–42

Sebastião R, Silva M, Rabiço R, Gama J, Mendonça T (2013) Real-time algorithm for changes detection in depth of anesthesia signals. Evolving Systems 4(1):3–12

Shankaranarayanan G, Cai Y (2006) Supporting data quality management in decision-making. Decision Support Systems 42(1):302–317

Shannon CE (1948) A mathematical theory of communication. The Bell System Technical Journal 27(1):379–423, 623–656

Shannon CE (2001) A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review 5(1):3–55

Shearer C (2000) The CRISP-DM Model: The New Blueprint for Data. Journal of Data Warehousing 5(4):13–22

Shewhart WA, Deming WE (1939) Statistical method from the viewpoint of quality control. Washington, D.C. : Graduate School of the Department of Agriculture

Shewhart WA, Deming WE (1986) Statistical method from the viewpoint of quality control. Dover

Shimazaki H, Shinomoto S (2007) A method for selecting the bin size of a time histogram. Neural Comput 19(6):1503–1527

Shimazaki H, Shinomoto S (2010) Kernel bandwidth optimization in spike rate estimation. J Comput Neurosci 29(1-2):171–182

Silverman B (1986) Density Estimation for Statistics and Data Analysis. Springer US

Solberg LI, Engebretson KI, Sperl-Hillen JM, Hroscikoski MC, O'Connor PJ (2006) Are claims data accurate enough to identify patients for performance measures or quality improvement? the case of diabetes, heart disease, and depression. American Journal of Medical Quality 21(4):238–245

Spiliopoulou M, Ntoutsi I, Theodoridis Y, Schult R (2006) Monic: Modeling and monitoring cluster transitions. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '06, pp 706–711

Österreicher F, Vajda I (2003) A new class of metric divergences on probability spaces and its applicability in statistics. Annals of the Institute of Statistical Mathematics 55(3):639–653

Stevens SS (1951) Mathematics, measurement, and psychophysics. Wiley

Stiglic G, Kokol P (2011) Interpretability of sudden concept drift in medical informatics domain. 2010 IEEE International Conference on Data Mining Workshops 0:609–613

Sullivan GM, Feinn R (2012) Using effect size—or why the *P* value is not enough. Journal of Graduate Medical Education 4(3):279–282

Sun K, Marchand-Maillet S (2014) An information geometry of statistical manifold learning. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp 1–9

Svolba G, Bauer P (1999) Statistical quality control in clinical trials. Controlled clinical trials 20(6):519–530

Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323

Torgerson W (1952) Multidimensional scaling: I. theory and method. Psychometrika 17(4):401–419

Tortajada S, Fuster-Garcia E, Vicente J, Wesseling P, Howe FA, Julià-Sapé M, Candiota AP, Monleón D, Moreno-Torres n, Pujol J, Griffiths JR, Wright A, Peet AC, Martínez-Bisbal MC, Celda B, Arús C, Robles M, García-Gómez JM (2011) Incremental gaussian discriminant analysis based on graybill and deal weighted combination of estimators for brain tumour diagnosis. Journal of Biomedical Informatics 44(4):677–687

Toubiana L, Cuggia M (2014) Big data and smart health strategies: Findings from the health information systems perspective:. IMIA Yearbook 9(1):125–127

Ullah A (1996) Entropy, divergence and distance measures with econometric applications. Journal of Statistical Planning and Inference 49(1):137–162

Verwey NA, van der Flier WM, Blennow K, Clark C, Sokolow S, De Deyn PP, Galasko D, Hampel H, Hartmann T, Kapaki E, Lannfelt L, Mehta PD, Parnetti L, Petzold A, Pirttila T, Saleh L, Skinningsrud A, Swieten JCv, Verbeek MM, Wiltfang J, Younkin S, Scheltens P, Blankenstein MA (2009) A worldwide multicentre comparison of assays for cerebrospinal fluid biomarkers in alzheimer's disease. Annals of Clinical Biochemistry 46(3):235–240

Vicente J, Sáez C, Tortajada S, Fuster E, García-Gómez JM (2012) Curiam BT kids, a clinical DSS for pediatric brain tumour diagnosis. vol 25, p 622

Wagnerova D, Herynek V, Malucelli A, Dezortova M, Vymazal J, Urgosik D, Syrucek M, Jiru F, Skoch A, Bartos R, Sames M, Hajek M (2012) Quantitative MR imaging and spectroscopy of brain tumours: a step forward? Eur Radiol 22(11):2307–2318

Walker KL, Kirillova O, Gillespie SE, Hsiao D, Pishchalenko V, Pai AK, Puro JE, Plumley R, Kudyakov R, Hu W, Allisany A, McBurnie M, Kurtz SE, Hazlehurst BL (2014) Using the CER hub to ensure data quality in a multi-institution smoking cessation study. Journal of the American Medical Informatics Association 21(6):1129–1135

Wang RY (1998) A product perspective on total data quality management. Commun ACM 41(2):58–65

Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. J Manage Inf Syst 12(4):5–33

Weber GM (2013) Federated queries of clinical data repositories: the sum of the parts does not equal the whole. J Am Med Inform Assoc 20(e1):e155–161

Weiskopf NG, Weng C (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 20(1):144–151

Wellings K, Macdowall W, Catchpole M, Goodrich J (1999) Seasonal variations in sexual activity and their implications for sexual health promotion. J R Soc Med 92(2):60–64

Westgard JO, Barry PL (2010) Basic QC practices: training in statistical quality control for medical laboratories. Westgard QC, Madison, WI

Whitney H (1940) Differentiable manifolds. Annals of Math 41:645–680

WHO (2009) World Health Organization. International statistical classification of diseases and related health problems. - 10th revision, 2008 edition. WHO Press

WHO (2012) World Health Organization. Strengthening civil registration and vital statistics for births, deaths and causes of death: resource kit. WHO Press

WHO, UNICEF (2009) World Health Organization and UNICEF. Baby-friendly hospital initiative: revised, updated and expanded for integrated care.

Widmer G, Kubat M (1996) Learning in the presence of concept drift and hidden contexts. Machine Learning 23(1):69–101

Wong PC, Foote H, Leung R, Adams D, Thomas J (2000) Data signatures and visualization of scientific data sets. Computer Graphics and Applications, IEEE 20(2):12–15

Xiao L, Peet A, Lewis P, Dashmapatra S, Sáez C, Croitoru M, Vicente J, Gonzalez-Velez H, Ariet MLi (2007) An adaptive security model for multi-agent systems and application to a clinical trials environment. IEEE, pp 261–268

Xiao L, Vicente J, Sáez C, Peet A, Gibb A, Lewis P, Dasmahapatra S, Croitoru M, Gonz H, Ariet M, Dupplaw D (2008) A security model and its application to a distributed decision support system for healthcare. IEEE, pp 578–585

Zhou A, Cai Z, Wei L (2015) Density estimation over data stream

Zhou SK, Chellappa R (2006) From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space. IEEE Trans Pattern Anal Mach Intell 28(6):917–929

Zurriaga Ó, Vanaclocha H, Martinez-Beneito MA, Botella-Rocamora P (2008) Spatio-temporal evolution of female lung cancer mortality in a region of spain, is it worth taking migration into account? BMC Cancer 8(1):35

Zurriaga Ó, López Briones C, Martínez-Beneito MA, Cavero-Carbonell C, Amorós R, Signes JM, Amador A, Sáez C, Robles M, García-Gómez JM, Navarro-Sánchez C, Sánchez-Pérez MJ, Vives-Corrons JL, Mañú MM, Olaya L (2015) Chapter 8: Running a registry. In: Zaletel M, Kralj M (eds) Methodological guidelines and recommendations for efficient and rational governance of patient registries, National Institute of Public Health, Trubarjeva 2, 1000 Ljubljana, Slovenia

Appendix A

Fisher Information Matrix

Let $p(x|\boldsymbol{\Theta})$ be the probability density function of a variable $x \in X$, uniquely parametrized by the vector parameter $\boldsymbol{\Theta}$ of a given family of probability distributions.

The log-likelihood function of p is defined by

$$\ell(X; \boldsymbol{\Theta}) = \log \mathcal{L}(X; \boldsymbol{\Theta}) = \log \prod_i p(x_i|\boldsymbol{\Theta}) \tag{A.1}$$

which represents the degree of adjustment of the value of $\boldsymbol{\Theta}$ w.r.t. the observed data in X.

The partial derivative (or gradient) of ℓ w.r.t. $\boldsymbol{\Theta}$ is known as the Fisher score or simply score:

$$s(X; \boldsymbol{\Theta}) = \nabla_{\boldsymbol{\Theta}} \ell(X; \boldsymbol{\Theta}) = \left(\frac{\partial}{\partial \Theta_1} \ell(X; \boldsymbol{\Theta}), ..., \frac{\partial}{\partial \Theta_N} \ell(X; \boldsymbol{\Theta}) \right) \tag{A.2}$$

which measures the sensitivity of ℓ in a sample X to changes in the values of the vector parameter $\boldsymbol{\Theta}$. Maximum Likelihood estimation attempts to estimate $\hat{\boldsymbol{\Theta}}$ as the true value of $\boldsymbol{\Theta}$ by means of finding a score equal to 0 (usually in a L2 norm).

For a fixed $\boldsymbol{\Theta}_j$, for all the possible samples X, and under some regularity conditions, the expected value of the score is 0:

$$\mathbb{E}[\nabla_{\boldsymbol{\Theta}} \ell(X; \boldsymbol{\Theta})] = 0 \tag{A.3}$$

On the other hand, a variance of the score $var(s(X; \boldsymbol{\Theta}_j))$ near to 0 means that most of the samples in X contain little information about the true value of $\boldsymbol{\Theta}$. That is, there are practically no regions in the space of X which adjust $\boldsymbol{\Theta}$. Consequently, considering $\hat{\boldsymbol{\Theta}}$ an unbiased estimator of $\boldsymbol{\Theta}$, its variance will also be large across X. On the contrary, a large variance of the score means that there exist regions of X with large information about $\boldsymbol{\Theta}$ (with large values of the score which make the variance increase), thus reducing the variance of its estimator $\hat{\boldsymbol{\Theta}}$. Such variance of the score is known as Fisher Information and, after some mathematical development assuming (A.3), is defined as:

$$\mathbb{I}(\boldsymbol{\Theta}) = \mathbb{E}[(\nabla_{\boldsymbol{\Theta}} \ell(X; \boldsymbol{\Theta}))^2] \tag{A.4}$$

The Fisher Information measures the amount of information about $\boldsymbol{\Theta}$ that is present in x. In addition, according to the Cramér-Rao inequality, which we will not discuss here, the Fisher Information gives a lower bound to the variance of an unbiased estimator $\hat{\boldsymbol{\Theta}}$ from X.

When $\boldsymbol{\Theta} = [\Theta_1, \Theta_2, ..., \Theta_N]^T$, $\mathbb{I}$ is provided as an NxN symmetric matrix, known as the Fisher Information Matrix (FIM), where

$$FIM(\boldsymbol{\Theta})_{i,j} = \mathbb{E}\left[\frac{\partial \ell(X;\boldsymbol{\Theta})}{\partial \Theta_i}\frac{\partial \ell(X;\boldsymbol{\Theta})}{\partial \Theta_j}\right] \tag{A.5}$$

We should mention that two parameters Θ_i and Θ_j are orthogonal, that is their MLE estimates can be calculated independently, when their joint component $\mathbb{I}(\boldsymbol{\Theta})_{i,j}$ in the FIM is zero. As an example, the FIM of the Normal distribution:

$$\begin{array}{c|cc} & \mu & \sigma \\ \hline \mu & \frac{1}{\sigma^2} & 0 \\ \sigma & 0 & \frac{1}{(2\sigma^4)} \end{array} \tag{A.6}$$

indicates that the MLE estimates of μ and σ are independent. Additionally, their respective estimation variances are given by the corresponding diagonal elements in FIM^{-1} (Cramér-Rao inequality).

The FIM defines the metric tensor, known as Fisher Information Metric, used as inner product in the Riemannian manifold in an N-dimensional parameter space, which allows applying differential geometry calculus in such an probability space.

Appendix B

Develoment of equations of simplex properties

B.1 Development of Equation 4.2: $d_{1R}(D)$

In any D-dimensional simplex Δ^D, between any pair of its vertices and the centroid a triangle is constituted. The three segments composing such triangle can be defined as $\overline{CV_i}$, $\overline{CV_j}$ and $\overline{V_iV_j}$, where V_i and V_j correspond to two vertices of Δ^D and C to its centroid. The angle between segments $\overline{CV_i}$ and $\overline{CV_j}$ is defined as $\angle\, \overline{CV_i}\ \overline{CV_j} = \gamma$.

Now, let Δ^D be a 1-regular simplex, Δ^D_{1R}, thus $\|\overline{V_iV_j}\| = 1$. Hence, γ, $\|\overline{CV_i}\|$ and $\|\overline{CV_j}\|$ will depend on D, and the following definitions apply:

1. $\|\overline{CV_i}\| = \|\overline{CV_j}\| = d_{1R}(D)$
2. $\gamma(D) = \arccos(\,^{-1}/_{D})$ (Parks and Wills, 2002)

Let the midpoint of $\overline{V_iV_j}$ be M. Hence, the median $\overline{CM}$ divides the triangle into two equal right-angled triangles, with $\|\overline{MV_i}\| = \|\overline{MV_j}\| = {}^1/_2$ and $\angle\, \overline{CM}\ \overline{CV_i} = \angle\, \overline{CM}\ \overline{CV_j} = {}^{\gamma}/_{2}$. Taking any of the two triangles, e.g., the one including the vertex V_i, according to its trigonometric functions:

$$\sin(^{\gamma}/_{2}) = \frac{\|\overline{MV_i}\|}{\|\overline{CV_i}\|}$$

As a consequence, replacing and solving the equation leads to:

$$d_{1R}(D) = \frac{1}{2\sin(^{\gamma(D)}/_{2})}$$

B.2 Development of Equation 4.3: $d_{max}(D)$

The centroid of any D-dimensional simplex, Δ^D, is calculated as:

$$C = \sum_{i=1}^{N} \frac{V_i}{N},$$

where V_i are the coordinates of vertex i and $N = D + 1$.

Let the distance between any two vertices $\overline{V_i V_j} \in [0, 1]$. For a simplex Δ^D, when $V_1 = V_2 = ... = V_{N-1}$, and $\|\overline{V_1 V_N}\| = 1$, the length of the segment $\overline{V_N C}$ will be maximum. Hence, corresponding to the distance of V_N to centroid, $\|\overline{V_N C}\| = d_{max}(D)$, and depends on the number of dimensions. In that situation, the centroid is calculated as:

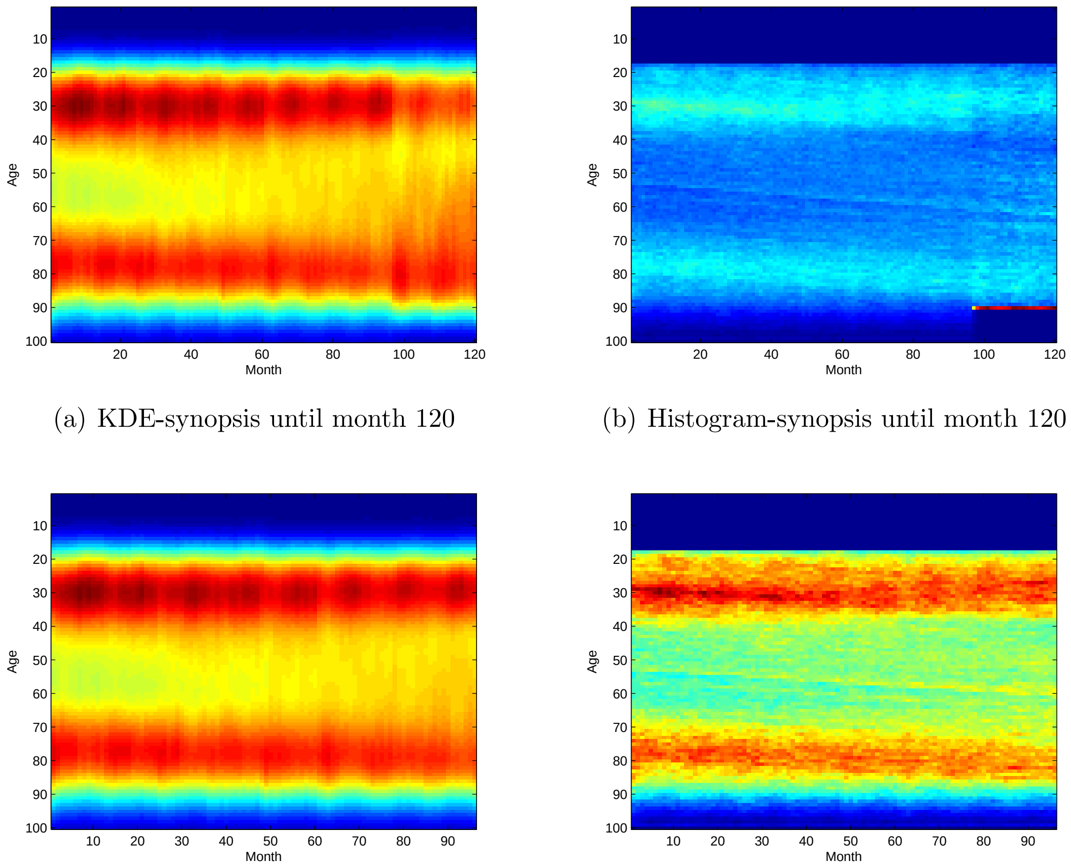
$$C = \frac{V_1 \cdot (N-1) + V_N}{N}$$

Let assume $V_1 = O$. By the conditions above $\|V_N\| = 1$. Hence,

$$\begin{aligned}
C &= \frac{V_N}{N} \\
C - V_N &= \frac{V_N}{N} - C + C - V_N && \text{Add } (C - V_N) \text{ to both sides} \\
-(C - V_N) &= V_N - \frac{V_N}{N} && \text{Multiply by } -1 \text{ both sides} \\
d_{max}(D) &= \|V_N\| - \frac{\|V_N\|}{N} && d_{max}(D) = \| - (C - V_N)\| \\
d_{max}(D) &= 1 - \frac{1}{N} && \|V_N\| = 1 \\
d_{max}(D) &= 1 - \frac{1}{D+1} && N = D + 1
\end{aligned}$$

Appendix C

Supplemental material for Chapter 5



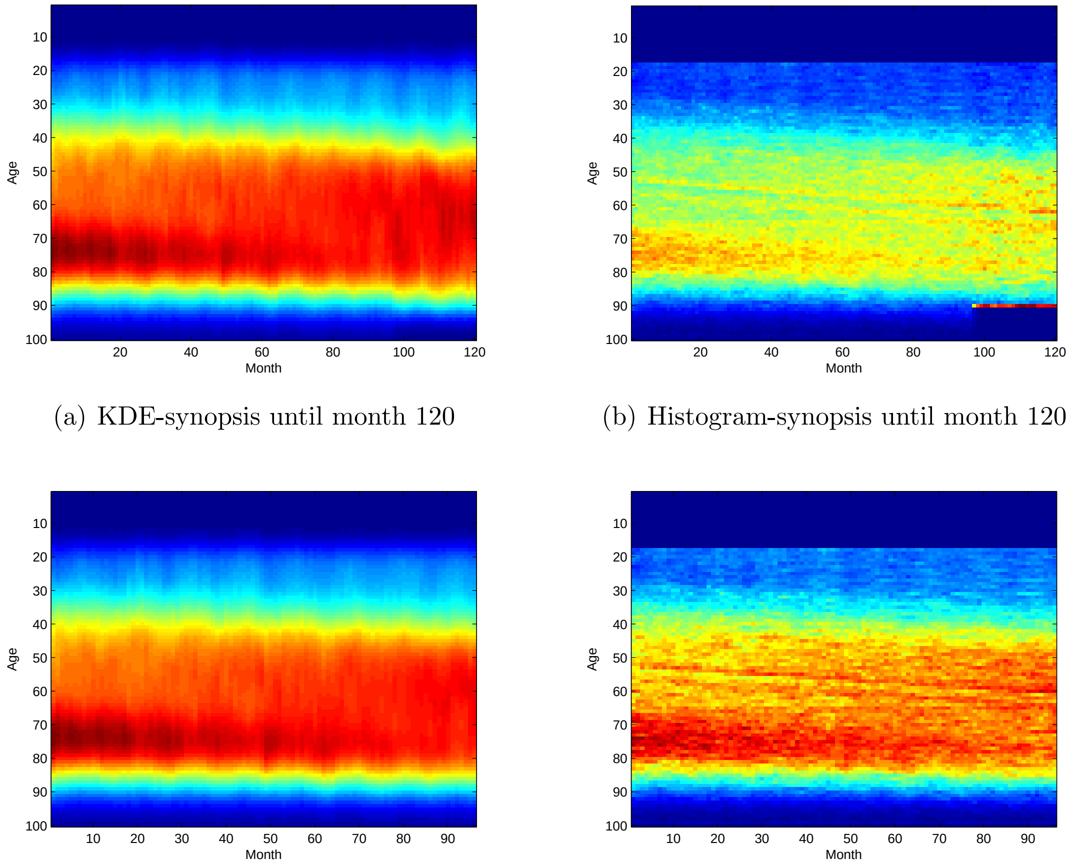
(a) KDE-synopsis until month 120

(b) Histogram-synopsis until month 120

(c) KDE-synopsis until month 96

(d) Histogram-synopsis until month 96

Figure C.1: Probability mass temporal maps of age on female patients from the NHDS dataset.



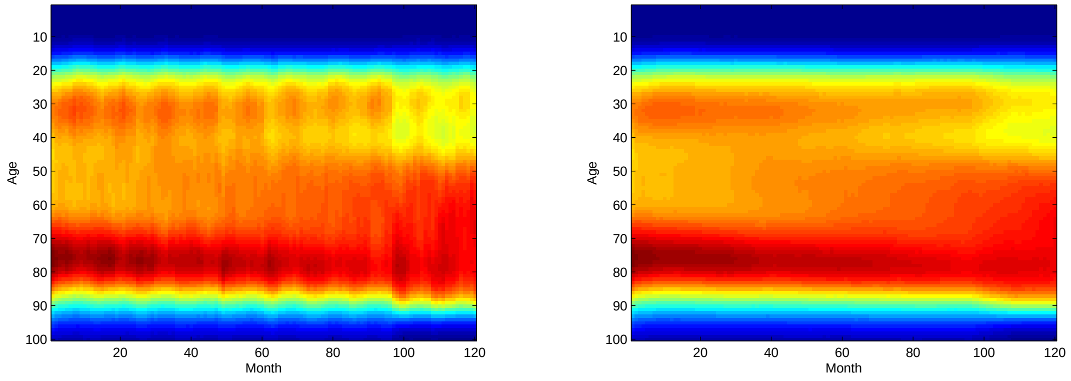
(a) KDE-synopsis until month 120

(b) Histogram-synopsis until month 120

(c) KDE-synopsis until month 96

(d) Histogram-synopsis until month 96

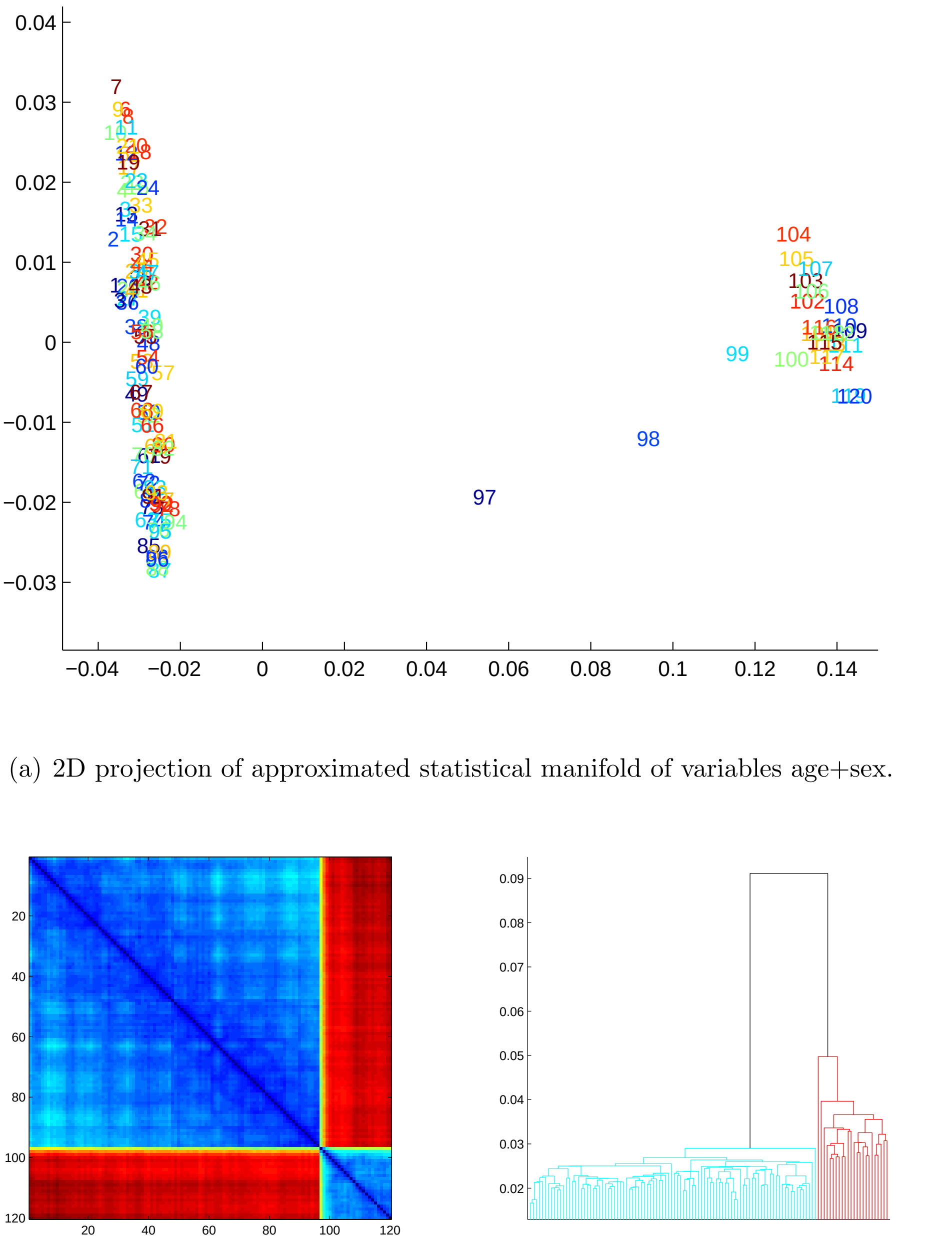
Figure C.2: Probability mass temporal maps of age on male patients from the NHDS dataset.



(a) Memoryless fading windows

(b) Sliding window ($w = 12$)

Figure C.3: Comparison of memoryless fading windows vs. sliding window with $w = 12$ approaches. The sliding window approach removes the recurrent effect while missing short-term information and delaying abrupt changes.



(a) 2D projection of approximated statistical manifold of variables age+sex.

(b) Dissimilarity matrix heatmap

(c) Dendrogram for temporal subgroups

Figure C.4: Change characterization and temporal subgroup discovery on the joint age and sex variables.

Appendix D

Basic examples of the variability methods

D.1 Multi-source variability

This section describes an intuitive example which summarizes the multi-source variability method. The example is described for three data sources (e.g., three sites or hospitals) where A, B and C represent the distributions of a variable under study on these data sources. Then, we define A, B and C normally distributed, with different parameters for mean (μ) and standard deviation (σ), and different sample sizes (n), as shown in Figure D.1.

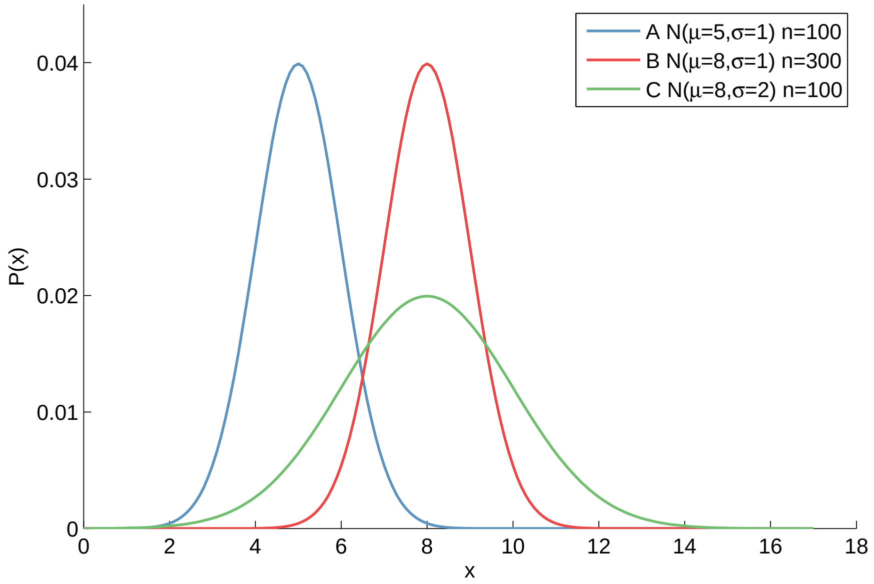


Figure D.1: Distributions of three simulated normally distributed variables representing three data sources

We next calculate the distances among the three distributions, concretely, the Jensen-Shannon distance.[1] We have then three distances, one for each pair of distributions. Based on these distances, we construct a geometric figure where points represent the distributions, and the distances among them are the previously calculated distances. In the case of three data sources, the geometric figure is a triangle, as shown in Figure D.2. In the general case, the figure is a simplex, the generalization of a triangle to multiple dimensions. In this simplex, the centroid represents a hidden average of all the distributions, and the distance of each distribution to it represents the Source Probabilistic Outlyingness (SPO) metric for each source.

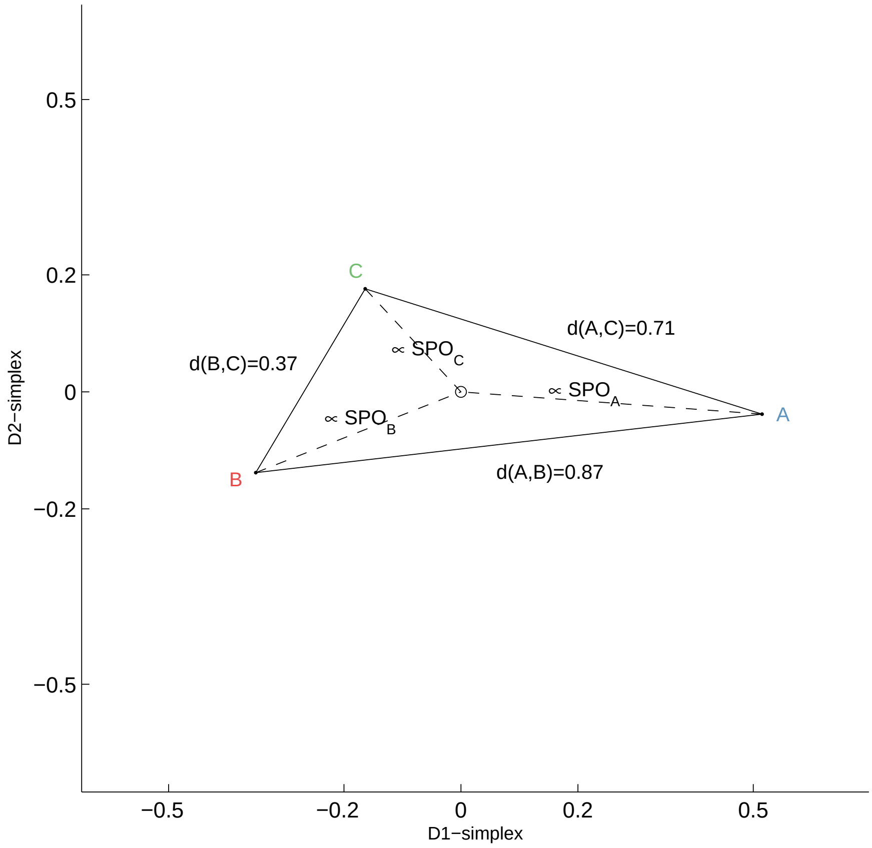


Figure D.2: Geometric figure constructed from the pairwise distances among the distributions A, B and C. The vertices represent the distributions, the edges the distances among them, and the centroid a hidden average distribution. The distances of the vertices to the centroid are proportional to the SPO metric for each source

The Global Probabilistic Deviation (GPD) is provided as the normalized mean of the distances of the vertices to the centroid. Then, it represents the standard deviation of all the data sources to the global average.

Finally, based on the calculated simplex we provide the multi-source simplex visualization in Figure D.3. In this visualization, each circle is located at the position of its associated data source in the previous simplex, then, the distances among them represent the dissimilarity among their distributions. Additionally, the circle color is associated to the source SPO metric, as the length to the hidden centroid in the simplex. Besides, the circle size is associated to the sample size of each source.

As a final remark, we note that in the case of more than three data sources, the corresponding simplex is a geometric figure represented in more than three dimensions (concretely in $D = N - 1$ dimensions, where N is the number of data sources). As a

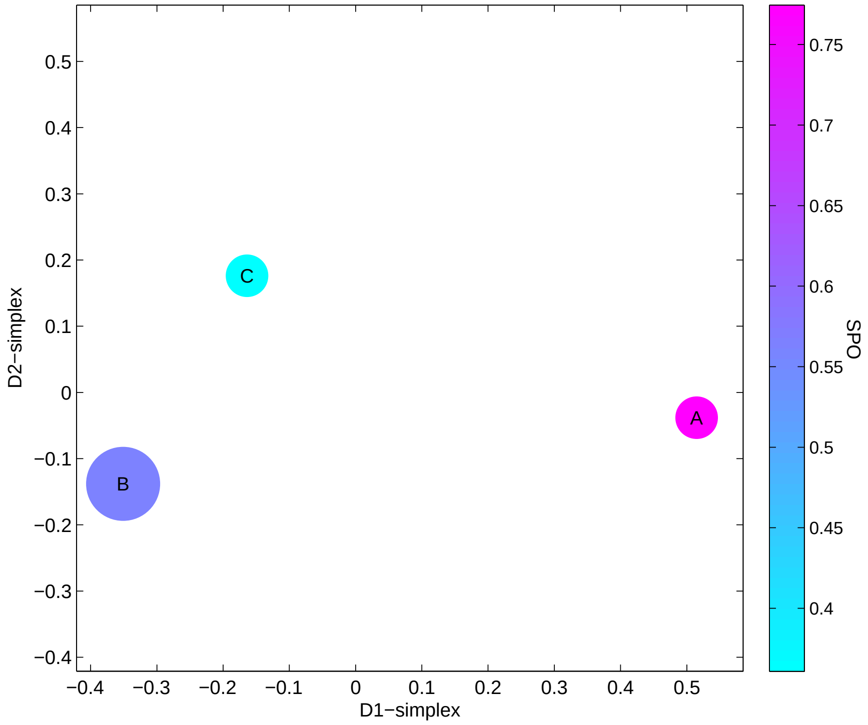


Figure D.3: Multi-source simplex visualization of the example. Each circle represents a data source, where the distances among them represent the distances among their distributions. The color indicates the SPO of each source, and the circle size the source sample size

consequence, it cannot be visualized in its original dimensionality. Using dimensionality reduction methods such as Multidimensional Scaling we can project such simplices into the two or three most representative dimensions (usually those with a larger variance), as shown in the multi-source simplex visualization. Therefore, we sometimes might note that the color of a circle shows a higher SPO than other data sources visually further than the former (respect to the centroid). This situation is normal when the distribution of a data source sums a larger distance to the centroid in dimensions further than the visualized; being such data source marginalized in the visualization by the larger variance among the other data sources.

D.2 Temporal variability

This section describes an intuitive example summarizing the temporal variability methods. The example is based on a simulated repository acquired from January 2014 to March 2015, which has been divided in 15 temporal batches, one per month. With the purpose of the example, the probability distribution of the repository data, a univariate normal distribution, has been varied through time in different manners (Figure D.4). First, from January 2014 to September 2014 the distribution mean have linearly and gradually moved. Second, in October 2014 the distribution mean have been abruptly moved respect to the previous month, and then continues gradually moving until December 2014. Finally, from January 2015 to March 2015, the mean remains fixed but the standard deviation is increased.

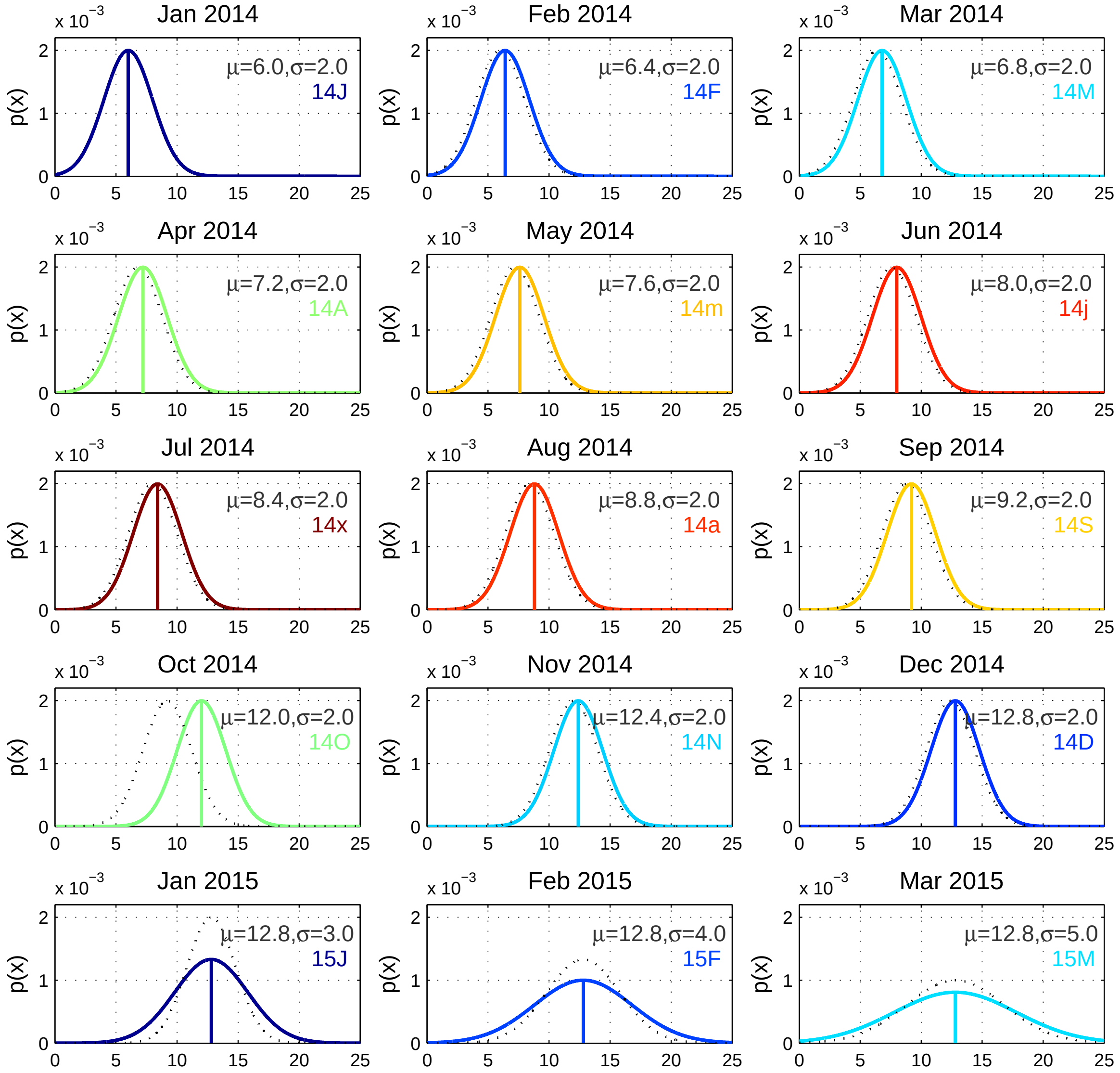


Figure D.4: Probality distributions of the 15 temporal batches, which temporal evolution can be appreciated. At right of each plot, the distribution mean (μ) and standard deviation (σ) are shown. For each plot, the dotted distribution shows the distribution of the immediately anterior temporal batch. The color and batch identifier at right are those that will be used in the IGT plot

The probability distribution temporal heat map of the repository simulated in this example is shown in Figure D.5. For each temporal batch (x axis) the heat map indicates with a color temperature the probability that a specific value (y axis) is observed. Basically, each column of this plot corresponds to the plot of the probability distribution at its associated temporal batch, as shown in Figure D.4. This heat map facilitates observing the changes in a single plot.

To construct the IGT plot we start by calculating the distances among the distributions of the 15 batches, similarly to the multi-source variability method, using the Jensen-Shannon distance. These distances can be organized in a matrix of 15 rows

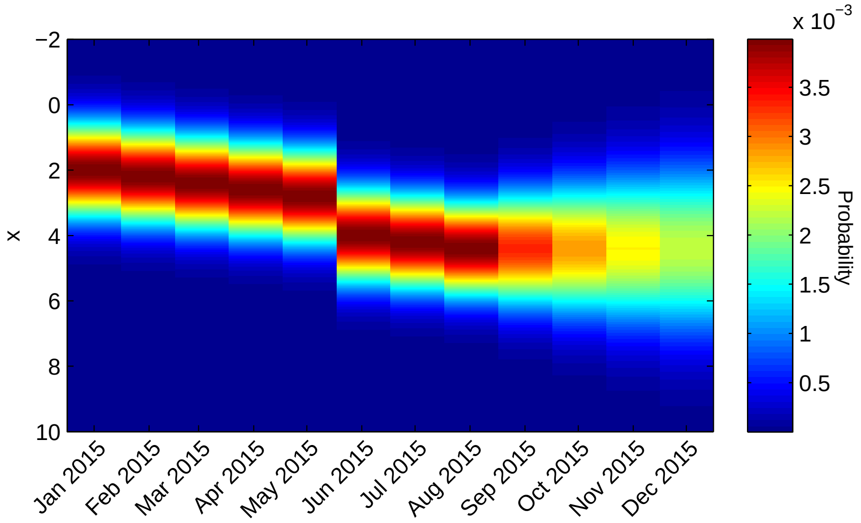


Figure D.5: Probability distribution temporal heatmap of the example

and 15 columns, where the value at a given index (i,j) represents the distance among distributions of temporal batches i and j. Based on this matrix, we can project the different temporal batches in a 2D plot using methods such as Multidimensional Scaling. Hence, temporal batches will be lied out as points in such a plot, and the distances among them will conserve the dissimilarities among their distributions (concretely, the dissimilarities shown in the plot are a 2D approximation of the original distances, in a higher dimensionality).

Figure D.6 shows the IGT plot resultant from the 15 temporal batches of the example. Hence, we can observe the gradual change from January 2014 (14J) to September 2014 (14S). Note that given the linear change in the distribution mean (increased by equal steps of 0.4 units, with a fixed standard deviation of 2, as seen in Figure D.4), contiguous months are equally spaced. Next we observe the abrupt change between September 2014 and October 2014 (14O), due to the abrupt change in the distribution mean (which increased suddenly increased in 2.8 units, as seen in Figure D.4). We next observe from October 2014 to December 2014 (14D) the same gradual increase as before. Finally, we observe a change of direction starting in January 2015 (15J) where, fixing the mean (in 12.8 units), the standard deviation began linearly increasing (in steps of 1 unit) until March 2015 (15M).

As shown, the IGT plot provides an exploratory visualization of the temporal behavior of data probability distributions. This can be supported with the quantitative measurement provided by the Probability Distribution Statistical Process Control (PDF-SPC) method. The PDF-SPC monitors an aggregated indicator of dissimilarity of distributions to a reference state. Hence, given a reference state, initially the first temporal batch, we measure the distance of consecutive distributions to that reference. These distances are bounded between zero and one, therefore the set of consecutive distances can be modelled by a Beta distribution. The difference of an upper confidence interval of the current Beta distribution to three reference confidence intervals (e.g., based on the three sigma rule) are used to classify the current degree of change in three

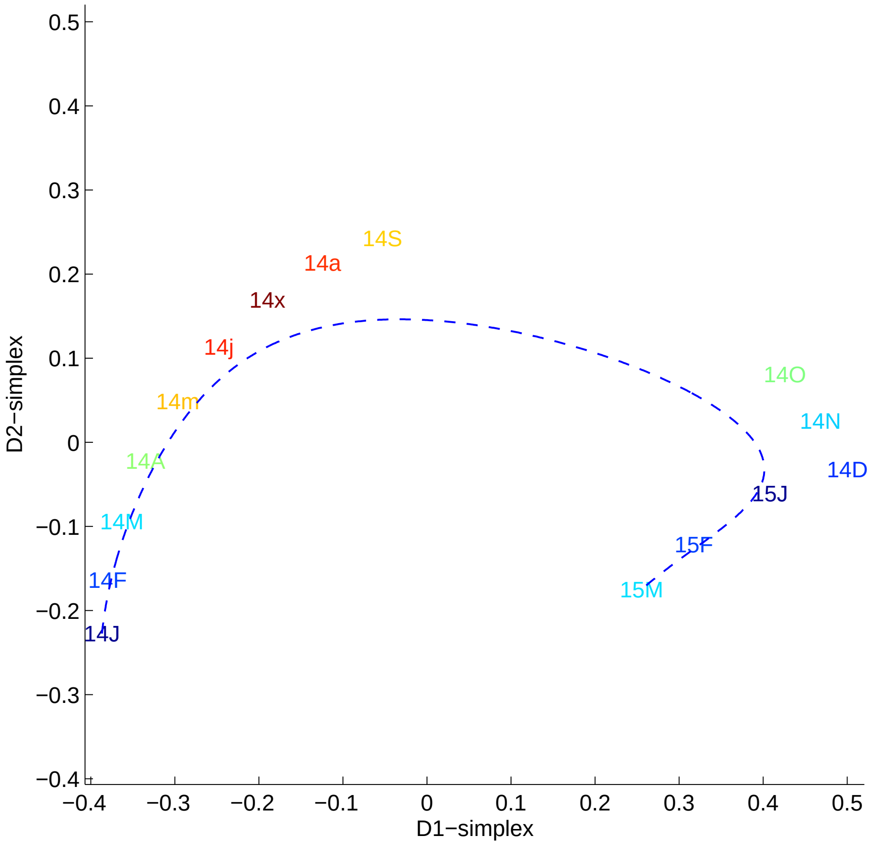


Figure D.6: IGT plot for the temporal batches of the example. Points represent the temporal batches, labelled with their dates in 'YYM' format, as the two latest digits of the year plus a single-character acronym for month with: J: January, F: February, M: March, A: April, m: May, j: June, x: July, a: August, S: September, O: October, N: November, D: December

states: in-control (distributions are stable), warning (distributions are changing), and out-of-control (recent distributions reached a significant dissimilarity to the reference leading to an unstable state). When an out-of-control state is reached, a significant change is confirmed and the reference distribution is set to the current.

Figure D.7, shows the resultant PDF-SPC monitoring from the 15 temporal batches of the example. After a transient state when the Beta distribution is stabilizing (and thus firing the expected false-alarm out-of-control state in March 2014), the simulated gradual change in distribution mean is detected as an increase of the monitored indicators, until a threshold is achieved in June 2014 firing a warning state. Next, in July 2014 a sufficiently large change is confirmed with an out-of-control state. Consequently, references are reestablished. But, next, due to the simulated abrupt change in October 2014, an out-of-control state is directly fired. Next, the following gradual changes in mean and standard deviation are captured as increases in the monitored indicators.

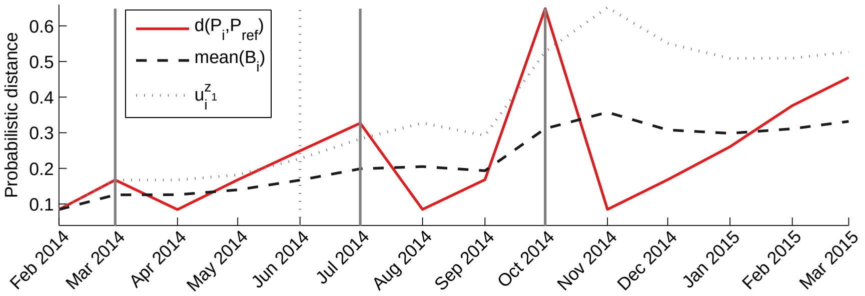


Figure D.7: PDF-SPC monitoring of the stability of the distribution of the example. The chart plots the current distance to the reference ($d(P_i, P_{ref})$), the mean of the accumulated distances ($\text{mean}(B_i)$), the upper confidence interval being monitored ($u_i^{z_1}$), and indicates the achievement of warning and out-of-control states as vertical dotted or continuous lines, respectively

Appendix E

Supplemental material for the Mortality Registry case study

E.1 WHO ICD-10 Mortality Condensed List 1

Table E.1: WHO ICD-10 Mortality Condensed List 1, excluding chapters

Code	Name	Code	Name
1-002	Cholera	1-050	Remainder of diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
1-003	Diarrhoea and gastroenteritis of presumed infectious origin	1-052	Diabetes mellitus
1-004	Other intestinal infectious diseases	1-053	Malnutrition
1-005	Respiratory tuberculosis	1-054	Remainder of endocrine, nutritional and metabolic diseases
1-006	Other tuberculosis	1-056	Mental and behavioural disorders due to pyschoactive substance use
1-007	Plague	1-057	Remainder of mental and behavioural disorders
1-008	Tetanus	1-059	Meningitis
1-009	Diphtheria	1-060	Alzheimer's disease
1-010	Whooping cough	1-061	Remainder of diseases of the nervous system
1-011	Meningococcal infection	1-062	Diseases of the eye and adnexa
1-012	Septicaemia	1-063	Diseases of the ear and mastoid process
1-013	Infections with a predominantly sexual mode of transmission	1-065	Acute rheumatic fever and chronic rheumatic heart diseases
1-014	Acute poliomyelitis	1-066	Hypertensive diseases
1-015	Rabies	1-067	Ischaemic heart diseases
1-016	Yellow fever	1-068	Other heart diseases
1-017	Other arthropod-borne viral fevers and viral haemorrhagic fevers	1-069	Cerebrovascular diseases
1-018	Measles	1-070	Atherosclerosis
1-019	Viral hepatitis	1-071	Remainder of diseases of the circulatory system
1-020	Human immunodeficiency virus [HIV] disease	1-073	Influenza
1-021	Malaria	1-074	Pneumonia
1-022	Leishmaniasis	1-075	Other acute lower respiratory infections
1-023	Trypanosomiasis	1-076	Chronic lower respiratory diseases
1-024	Schistosomiasis	1-077	Remainder of diseases of the respiratory system
1-025	Remainder of certain infectious and parasitic diseases	1-079	Gastric and duodenal ulcer
1-027	Malignant neoplasm of lip, oral cavity and pharynx	1-080	Diseases of the liver
1-028	Malignant neoplasm of oesophagus	1-081	Remainder of diseases of the digestive system
1-029	Malignant neoplasm of stomach	1-082	Diseases of the skin and subcutaneous tissue
1-030	Malignant neoplasm of colon, rectum and anus	1-083	Diseases of the musculoskeletal system and connective tissue
1-031	Malignant neoplasm of liver and intrahepatic bile ducts	1-085	Glomerular and renal tubulo-interstitial diseases
1-032	Malignant neoplasm of pancreas	1-086	Remainder of diseases of the genitourinary system
1-033	Malignant neoplasm of larynx	1-088	Pregnancy with abortive outcome
1-034	Malignant neoplasm of trachea, bronchus and lung	1-089	Other direct obstetric deaths
1-035	Malignant melanoma of skin	1-090	Indirect obstetric deaths
1-036	Malignant neoplasm of breast	1-091	Remainder of pregnancy, childbirth and the puerperium
1-037	Malignant neoplasm of cervix uteri	1-092	Certain conditions originating in the perinatal period
1-038	Malignant neoplasm of other and unspecified parts of uterus	1-093	Congenital malformations, deformations and chromosomal abnormalities
1-039	Malignant neoplasm of ovary	1-094	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
1-040	Malignant neoplasm of prostate	1-096	Transport accidents
1-041	Malignant neoplasm of bladder	1-097	Falls
1-042	Malignant neoplasm of meninges, brain and other parts of central nervous system	1-098	Accidental drowning and submersion
1-043	Non-Hodgkin's lymphoma	1-099	Exposure to smoke, fire and flames
1-044	Multiple myeloma and malignant plasma cell neoplasms	1-100	Accidental poisoning by and exposure to noxious substances
1-045	Leukaemias	1-101	Intentional self-harm
1-046	Remainder of malignant neoplasms	1-102	Assault
1-047	Remainder of neoplasms	1-103	All other external causes
1-049	Anaemias	1-901	SARS

E.2 Sample size tables

Table E.2: Sample sizes by Health Department in the Mortality Registry for both females and males (number of deaths)

	Year													
Department	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	**TOTAL**
AGral	301	326	292	357	331	372	365	367	403	429	399	443	419	**4804**
Requena	574	560	558	597	553	563	542	559	582	540	585	565	540	**7318**
SantJoan	485	533	546	640	590	614	595	615	598	565	648	644	663	**7736**
ClínicMr	747	660	775	731	741	713	666	761	739	776	777	757	843	**9686**
Vinaròs	748	776	763	855	804	786	789	773	845	854	792	824	836	**10445**
Peset	799	831	848	996	886	927	847	889	907	1025	914	927	1008	**11804**
Torrevieja	760	904	939	1009	987	1087	1078	1144	1171	1126	1115	1176	1294	**13790**
MarinaB	1089	1127	1096	1241	1232	1247	1276	1269	1282	1190	1250	1306	1449	**16054**
Alcoi	1190	1200	1261	1211	1193	1285	1167	1274	1266	1228	1283	1276	1402	**16236**
Orihuela	1028	1147	1179	1298	1290	1334	1272	1263	1260	1216	1261	1359	1373	**16280**
Sagunt	1195	1244	1210	1351	1258	1383	1259	1354	1312	1396	1353	1430	1437	**17182**
Manises	1333	1304	1368	1377	1367	1467	1345	1435	1431	1490	1419	1466	1532	**18334**
VGral	1253	1377	1331	1417	1323	1499	1481	1483	1414	1464	1469	1479	1513	**18503**
Dénia	1388	1334	1435	1464	1399	1432	1425	1461	1582	1434	1556	1531	1668	**19109**
Gandia	1388	1428	1456	1534	1432	1544	1425	1509	1525	1503	1494	1566	1639	**19443**
LaPlana	1422	1474	1496	1546	1485	1578	1538	1458	1564	1520	1528	1593	1524	**19726**
Elda	1414	1532	1419	1516	1517	1573	1427	1607	1537	1609	1601	1550	1795	**20097**
XàtivaOnt	1703	1665	1740	1961	1832	1871	1712	1816	1827	1854	1864	1846	1990	**23681**
Elx	1797	1790	1830	1961	1932	1914	1978	2003	2025	2136	2147	2080	2272	**25865**
ArnauLlíria	1932	1814	1939	1991	1919	2017	1962	2121	2098	2033	2097	2094	2171	**26188**
Castelló	2038	2104	2083	2206	2050	2246	2112	2176	2107	2222	2068	2110	2189	**27711**
LaRibera	2185	2246	2388	2361	2274	2392	2280	2395	2352	2371	2375	2446	2527	**30592**
Alacant	2274	2373	2319	2466	2363	2427	2382	2504	2504	2553	2491	2678	2671	**32005**
València	6928	6949	7133	7478	7164	7390	7130	7156	7145	7019	7071	7109	7066	**92738**
GLOBAL	**35971**	**36698**	**37404**	**39564**	**37922**	**39661**	**38053**	**39392**	**39476**	**39553**	**39557**	**40255**	**41821**	**505327**

Table E.3: Sample sizes by Health Department in the Mortality Registry for females (number of deaths)

	Year													
Department	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	**TOTAL**
Agral	143	148	142	176	155	185	156	174	199	184	179	209	183	**2233**
Requena	268	277	233	272	246	258	253	236	284	252	283	274	268	**3404**
SantJoan	219	272	266	317	298	275	298	285	287	274	312	318	321	**3742**
ClínicMr	355	310	357	330	363	340	311	379	367	402	357	377	384	**4632**
Vinaròs	371	361	367	415	386	366	368	347	401	386	363	400	383	**4914**
Torrevieja	295	355	370	420	373	444	459	456	475	440	474	479	528	**5568**
Peset	382	386	400	471	417	459	407	427	462	478	420	428	484	**5621**
MarinaB	475	508	473	569	550	594	587	548	547	564	570	598	639	**7222**
Orihuela	483	535	572	608	609	625	564	548	556	559	590	608	594	**7451**
Alcoi	595	558	608	607	587	616	547	641	621	635	652	631	688	**7986**
Sagunt	562	605	567	665	605	663	607	669	644	686	656	691	674	**8294**
Manises	614	640	646	651	629	680	627	637	702	685	662	721	741	**8635**
Dénia	612	596	630	709	626	663	636	656	694	642	720	685	781	**8650**
VGral	624	647	606	681	623	703	714	716	679	696	707	661	761	**8818**
Gandia	668	671	699	694	697	732	662	716	687	706	693	721	802	**9148**
LaPlana	680	696	738	737	687	745	708	698	760	732	750	754	740	**9425**
Elda	697	727	669	762	747	769	715	781	719	785	754	773	910	**9808**
XàtivaOnt	772	767	856	968	892	940	811	881	895	922	855	877	1005	**11441**
Elx	817	825	838	900	910	899	919	911	948	998	1040	979	1056	**12040**
ArnauLlíria	907	892	914	977	879	954	960	991	1031	989	1027	1017	1088	**12626**
Castelló	934	998	1009	1063	978	1048	972	1041	1055	1034	1011	1016	1037	**13196**
LaRibera	1046	1102	1117	1133	1104	1151	1090	1143	1147	1100	1163	1187	1189	**14672**
Alacant	1062	1106	1072	1141	1108	1159	1144	1202	1184	1241	1197	1345	1322	**15283**
València	3399	3427	3435	3664	3536	3740	3576	3650	3620	3513	3514	3605	3665	**46344**
GLOBAL	**16980**	**17409**	**17584**	**18930**	**18005**	**19008**	**18091**	**18733**	**18964**	**18903**	**18949**	**19354**	**20243**	**241153**

Table E.4: Sample sizes by Health Department in the Mortality Registry for males (number of deaths)

	Year													
Department	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	**TOTAL**
AGral	158	178	150	181	176	187	209	193	204	245	220	234	236	**2571**
Requena	306	283	325	325	307	305	289	323	298	288	302	291	272	**3914**
SantJoan	266	261	280	323	292	339	297	330	311	291	336	326	342	**3994**
ClínicMr	392	350	418	401	378	373	355	382	372	374	420	380	459	**5054**
Vinaròs	377	415	396	440	418	420	421	426	444	468	429	424	453	**5531**
Peset	417	445	448	525	469	468	440	462	445	547	494	499	524	**6183**
Torrevieja	465	549	569	589	614	643	619	688	696	686	641	697	766	**8222**
Alcoi	595	642	653	604	606	669	620	633	645	593	631	645	714	**8250**
Orihuela	545	612	607	690	681	709	708	715	704	657	671	751	779	**8829**
MarinaB	614	619	623	672	682	653	689	721	735	626	680	708	810	**8832**
Sagunt	633	639	643	686	653	720	652	685	668	710	697	739	763	**8888**
VGral	629	730	725	736	700	796	767	767	735	768	762	818	752	**9685**
Manises	719	664	722	726	738	787	718	798	729	805	757	745	791	**9699**
Elda	717	805	750	754	770	804	712	826	818	824	847	777	885	**10289**
Gandia	720	757	757	840	735	812	763	793	838	797	801	845	837	**10295**
LaPlana	742	778	758	809	798	833	830	760	804	788	778	839	784	**10301**
Dénia	776	738	805	755	773	769	789	805	888	792	836	846	887	**10459**
XàtivaOnt	931	898	884	993	940	931	901	935	932	932	1009	969	985	**12240**
ArnauLlíria	1025	922	1025	1014	1040	1063	1002	1130	1067	1044	1070	1077	1083	**13562**
Elx	980	965	992	1061	1022	1015	1059	1092	1077	1138	1107	1101	1216	**13825**
Castelló	1104	1106	1074	1143	1072	1198	1140	1135	1052	1188	1057	1094	1152	**14515**
LaRibera	1139	1144	1271	1228	1170	1241	1190	1252	1205	1271	1212	1259	1338	**15920**
Alacant	1212	1267	1247	1325	1255	1268	1238	1302	1320	1312	1294	1333	1349	**16722**
València	3529	3522	3698	3814	3628	3650	3554	3506	3525	3506	3557	3504	3401	**46394**
GLOBAL	**18991**	**19289**	**19820**	**20634**	**19917**	**20653**	**19962**	**20659**	**20512**	**20650**	**20608**	**20901**	**21578**	**264174**

E.3 Temporal heat maps of intermediate cause 1 and 2

This section shows the probability distribution temporal heat maps of the variables *IntermediateCause1* (figure E.1) and *IntermediateCause2* (figure E.2). Two main findings can be observed in both figures. The first is the punctual increment of unfilled data, labelled as NA, in January to March 2000, especially in February. The second is the abrupt change in frequencies of several causes (specially the NA), in March 2009.

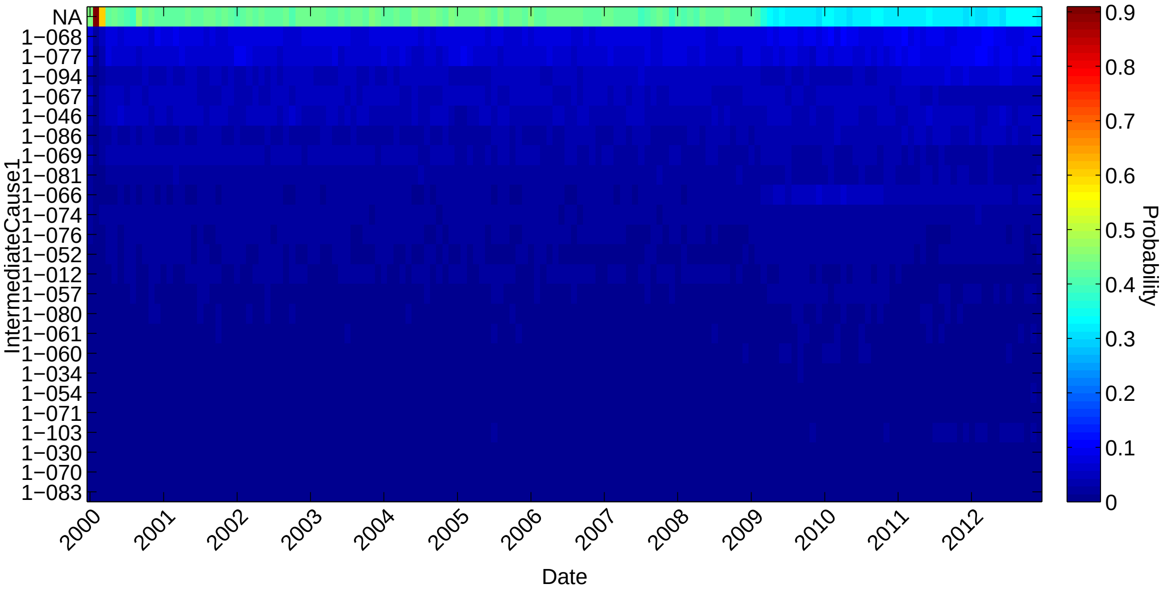


Figure E.1: PDF temporal heat map of *IntermediateCause1* for both sexes (see codes in Table E.1)

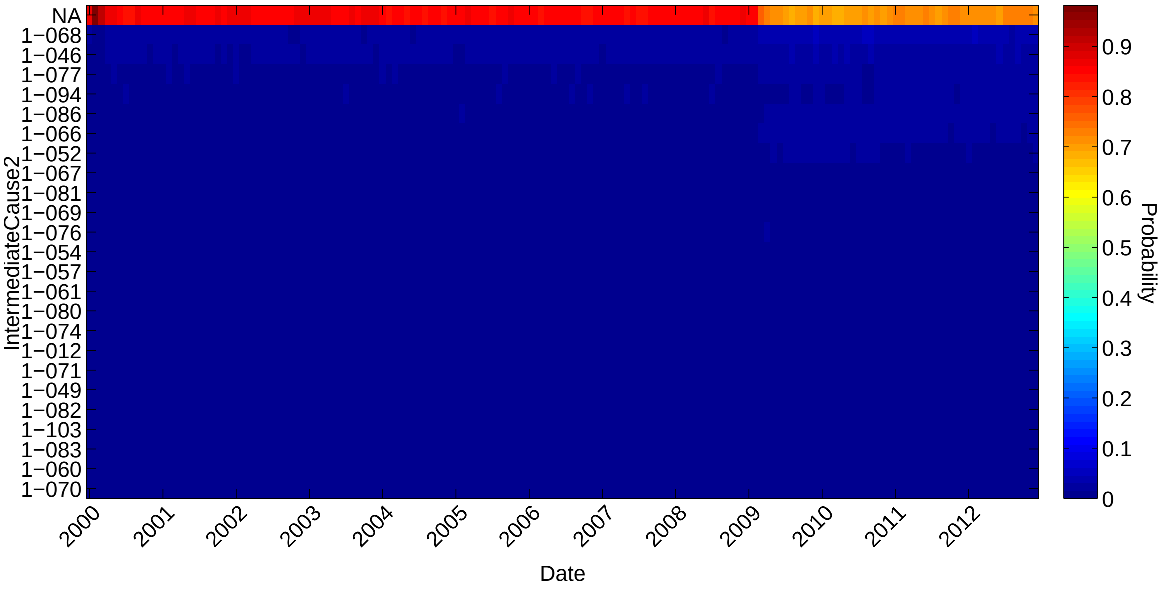


Figure E.2: PDF temporal heat map of *IntermediateCause2* for both sexes (see codes in Table E.1)

E.4 Unfilled values in the Death Certificate by Health Department

The heat maps in this section show the percentage of unfilled values by Health Department for each possible cause. To avoid the differences caused by the change of certificate in 2009, two heatmaps are shown with the data before (Figure E.3) and after (Figure E.4) such change. The Health Departments are sorted by the total number of unfilled causes.

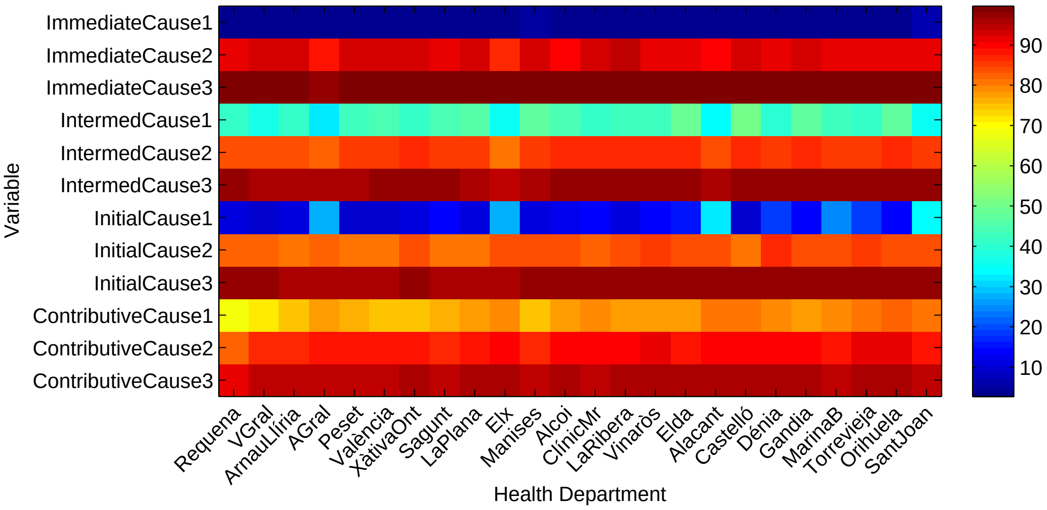


Figure E.3: Percentage of unfilled values by Health Department before the change of Certificate of Death (period 2000 - 2009 February)

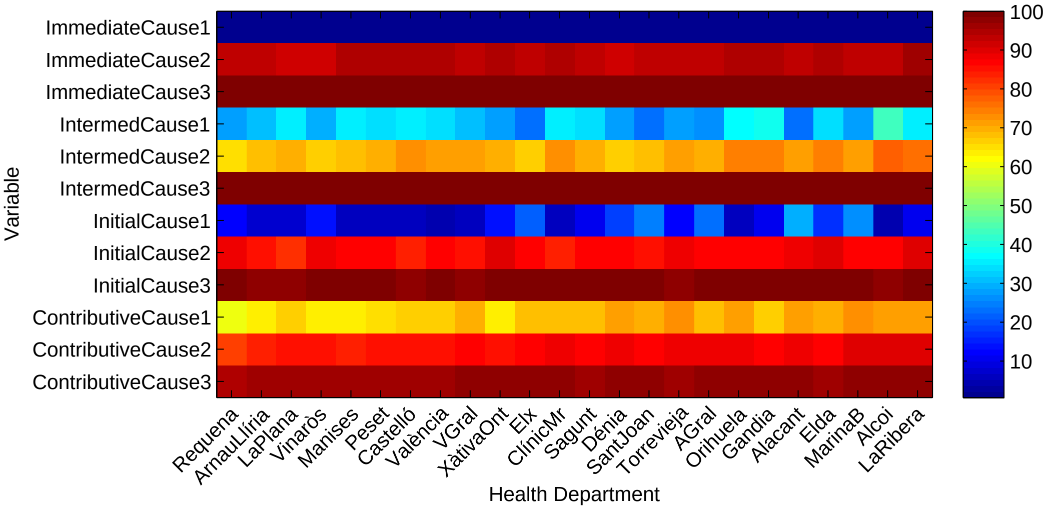


Figure E.4: Percentage of unfilled values by Health Department after the change of Certificate of Death (period 2009 March - 2012)

E.5 Multi-site variability of age of death

The figures in this section show the multi-source variability simplices of the Age of death in males (Figure E.5) and females (Figure E.6) during all the period of study. The most remarkable finding is the extreme outlyingness of the Department of Torrevieja in both sexes. Torrevieja counts with a large number of deaths in young people, in comparison with the rest. Additionally, in the case of males it can be appreciated that it is opposite to Requena, one of the Departments with an elder population.

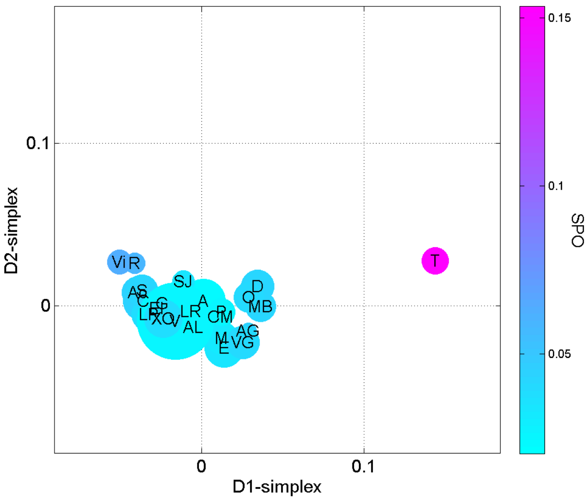


Figure E.5: Multi-site 2D simplices for Age in females during all the period of study

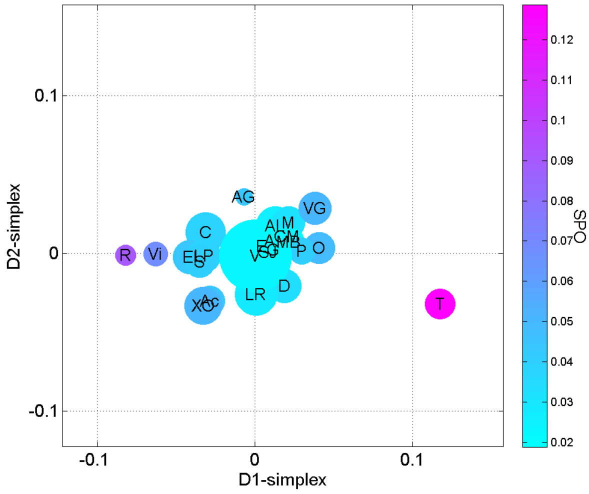


Figure E.6: Multi-site 2D simplices for Age in males during all the period of study

E.6 Dendrograms of initial cause 1 and intermediate cause 2

This section show the resultant dendrograms (concretely, phylogenetic trees) of the clustering process applied to the Health Departments in the variables *InitialCause1* (Figure E.7) and *IntermediateCause1* (Figure E.8) in males. The clustering of multi-source distributions can be performed based on the dissimilarity matrix of inter-source probabilistic distances resultant of the temporal stability method.

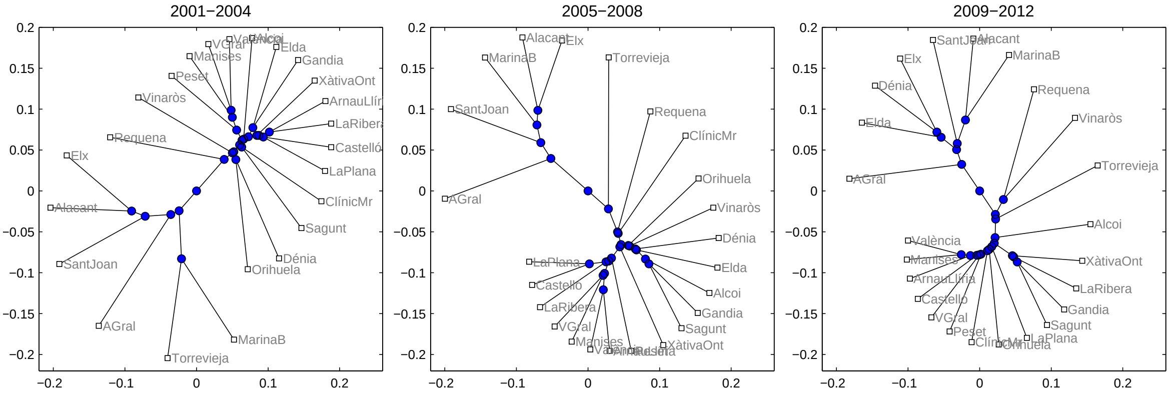


Figure E.7: Resultant dendrograms from clustering of InitialCause1 by Health Departments

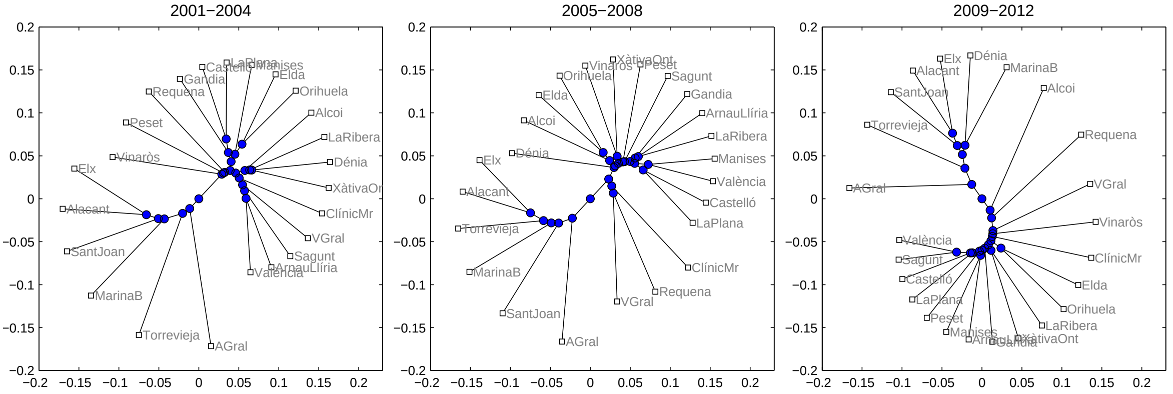
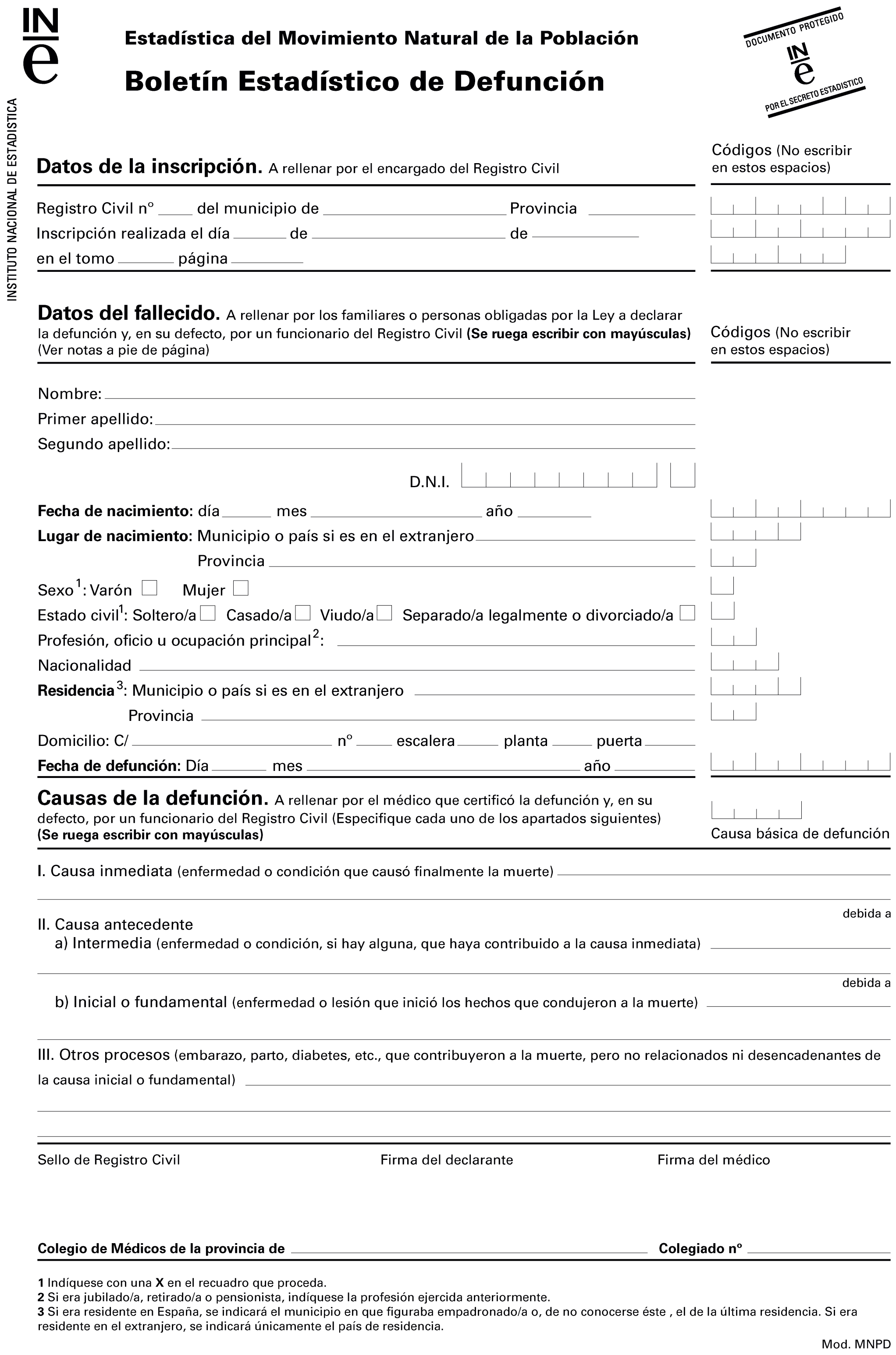


Figure E.8: Resultant dendrograms from clustering of IntermediateCause1 by Health Departments

E.7 Spanish Certificates of Death in the period 2000-2012



INSTITUTO NACIONAL DE ESTADISTICA

Estadística del Movimiento Natural de la Población

Boletín Estadístico de Defunción

DOCUMENTO PROTEGIDO POR EL SECRETO ESTADISTICO

Datos de la inscripción. A rellenar por el encargado del Registro Civil

Códigos (No escribir en estos espacios)

Registro Civil nº ____ del municipio de ____ Provincia ____

Inscripción realizada el día ____ de ____ de ____

en el tomo ____ página ____

Datos del fallecido. A rellenar por los familiares o personas obligadas por la Ley a declarar la defunción y, en su defecto, por un funcionario del Registro Civil **(Se ruega escribir con mayúsculas)** (Ver notas a pie de página)

Códigos (No escribir en estos espacios)

Nombre: ____

Primer apellido: ____

Segundo apellido: ____

D.N.I.

Fecha de nacimiento: día ____ mes ____ año ____

Lugar de nacimiento: Municipio o país si es en el extranjero ____

Provincia ____

Sexo[1]: Varón □ Mujer □

Estado civil[1]: Soltero/a □ Casado/a □ Viudo/a □ Separado/a legalmente o divorciado/a □

Profesión, oficio u ocupación principal[2]: ____

Nacionalidad ____

Residencia[3]: Municipio o país si es en el extranjero ____

Provincia ____

Domicilio: C/ ____ nº ____ escalera ____ planta ____ puerta ____

Fecha de defunción: Día ____ mes ____ año ____

Causas de la defunción. A rellenar por el médico que certificó la defunción y, en su defecto, por un funcionario del Registro Civil (Especifique cada uno de los apartados siguientes) **(Se ruega escribir con mayúsculas)**

Causa básica de defunción

I. Causa inmediata (enfermedad o condición que causó finalmente la muerte) ____

debida a

II. Causa antecedente

a) Intermedia (enfermedad o condición, si hay alguna, que haya contribuido a la causa inmediata) ____

debida a

b) Inicial o fundamental (enfermedad o lesión que inició los hechos que condujeron a la muerte) ____

III. Otros procesos (embarazo, parto, diabetes, etc., que contribuyeron a la muerte, pero no relacionados ni desencadenantes de la causa inicial o fundamental) ____

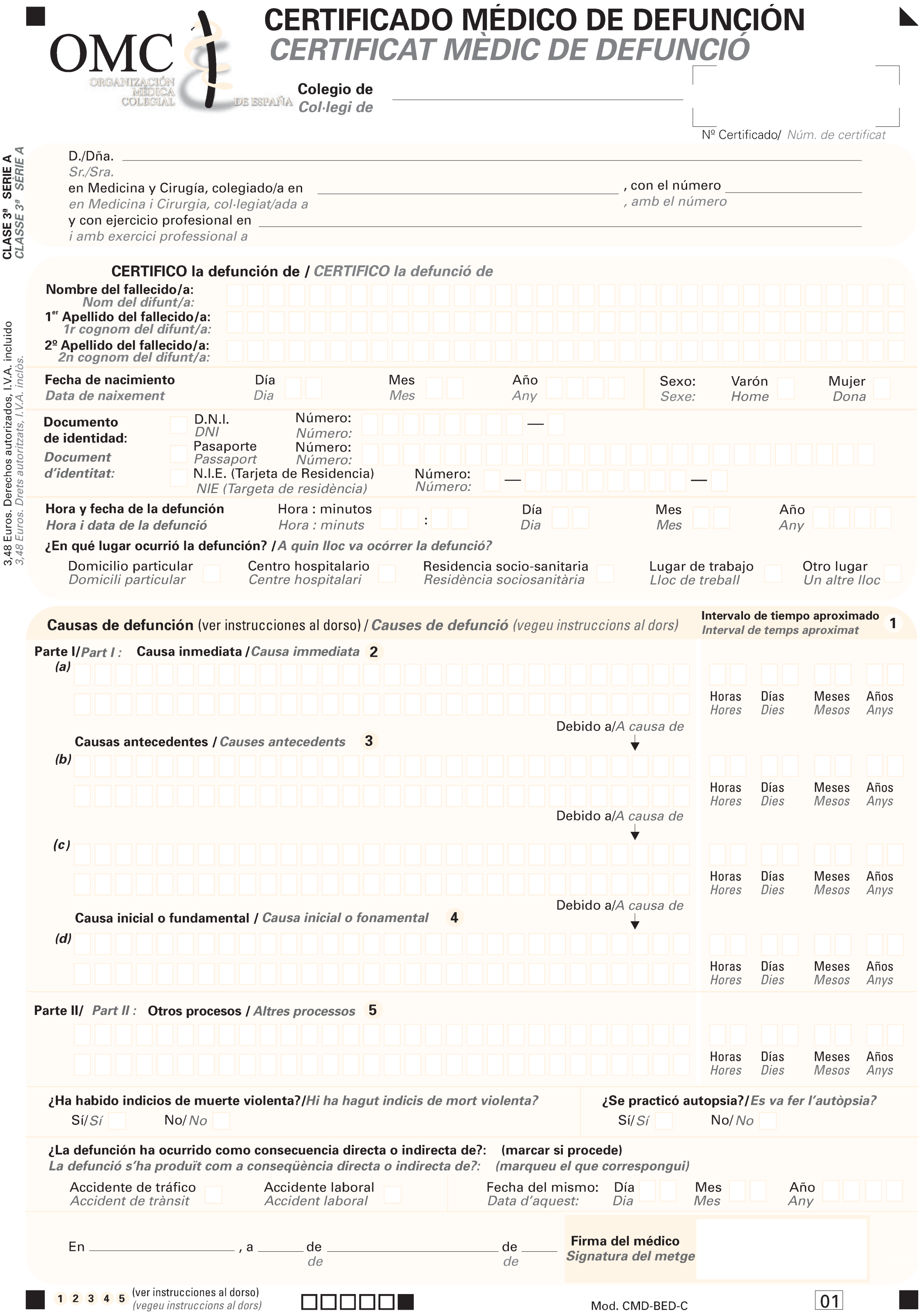
Sello de Registro Civil | Firma del declarante | Firma del médico

Colegio de Médicos de la provincia de ____ **Colegiado nº** ____

1 Indíquese con una **X** en el recuadro que proceda.
2 Si era jubilado/a, retirado/a o pensionista, indíquese la profesión ejercida anteriormente.
3 Si era residente en España, se indicará el municipio en que figuraba empadronado/a o, de no conocerse éste , el de la última residencia. Si era residente en el extranjero, se indicará únicamente el país de residencia.

Mod. MNPD

Figure E.9: Boletín Estadístico de Defunción (version 1999)



OMC ORGANIZACIÓN MÉDICA COLEGIAL DE ESPAÑA

CERTIFICADO MÉDICO DE DEFUNCIÓN
CERTIFICAT MÈDIC DE DEFUNCIÓ

Colegio de / *Col·legi de* ____________________

Nº Certificado/ *Núm. de certificat*

CLASE 3ª SERIE A / *CLASSE 3ª SERIE A*

3,48 Euros. Derechos autorizados, I.V.A. incluido / *3,48 Euros. Drets autoritzats, I.V.A. inclòs.*

D./Dña. / *Sr./Sra.* ____________________

en Medicina y Cirugía, colegiado/a en / *en Medicina i Cirurgia, col·legiat/ada a* ____________________ , con el número / *, amb el número* ____________

y con ejercicio profesional en / *i amb exercici professional a* ____________________

CERTIFICO la defunción de / *CERTIFICO la defunció de*

Nombre del fallecido/a: / *Nom del difunt/a:*

1er Apellido del fallecido/a: / *1r cognom del difunt/a:*

2º Apellido del fallecido/a: / *2n cognom del difunt/a:*

Fecha de nacimiento / *Data de naixement* — Día / *Dia* □□ Mes / *Mes* □□ Año / *Any* □□□□

Sexo: / *Sexe:* Varón / *Home* □ Mujer / *Dona* □

Documento de identidad: / *Document d'identitat:*

□ D.N.I. / *DNI* — Número: / *Número:* □□□□□□□□ — □

□ Pasaporte / *Passaport* — Número: / *Número:*

□ N.I.E. (Tarjeta de Residencia) / *NIE (Targeta de residència)* — Número: / *Número:* □ — □□□□□□□□ — □

Hora y fecha de la defunción / *Hora i data de la defunció* — Hora : minutos / *Hora : minuts* □□ : □□ Día / *Dia* □□ Mes / *Mes* □□ Año / *Any* □□□□

¿En qué lugar ocurrió la defunción? / *A quin lloc va ocórrer la defunció?*

Domicilio particular / *Domicili particular* □ Centro hospitalario / *Centre hospitalari* □ Residencia socio-sanitaria / *Residència sociosanitària* □ Lugar de trabajo / *Lloc de treball* □ Otro lugar / *Un altre lloc* □

Causas de defunción (ver instrucciones al dorso) / ***Causes de defunció*** *(vegeu instruccions al dors)*

Intervalo de tiempo aproximado / ***Interval de temps aproximat*** 1

Parte I/ *Part I :* **Causa inmediata /** *Causa immediata* 2

(a) — Horas / *Hores* Días / *Dies* Meses / *Mesos* Años / *Anys*

Debido a/*A causa de*

Causas antecedentes / *Causes antecedents* 3

(b) — Horas / *Hores* Días / *Dies* Meses / *Mesos* Años / *Anys*

Debido a/*A causa de*

(c) — Horas / *Hores* Días / *Dies* Meses / *Mesos* Años / *Anys*

Debido a/*A causa de*

Causa inicial o fundamental / *Causa inicial o fonamental* 4

(d) — Horas / *Hores* Días / *Dies* Meses / *Mesos* Años / *Anys*

Parte II/ *Part II :* **Otros procesos /** *Altres processos* 5

Horas / *Hores* Días / *Dies* Meses / *Mesos* Años / *Anys*

¿Ha habido indicios de muerte violenta?/*Hi ha hagut indicis de mort violenta?* Sí/*Sí* □ No/ *No* □

¿Se practicó autopsia?/*Es va fer l'autòpsia?* Sí/ *Sí* □ No/ *No* □

¿La defunción ha ocurrido como consecuencia directa o indirecta de?: (marcar si procede)
La defunció s'ha produït com a conseqüència directa o indirecta de?: (marqueu el que correspongui)

Accidente de tráfico / *Accident de trànsit* □ Accidente laboral / *Accident laboral* □ Fecha del mismo: / *Data d'aquest:* Día / *Dia* □□ Mes / *Mes* □□ Año / *Any* □□□□

En ____________ , a ______ de / *de* ____________ de / *de* ______

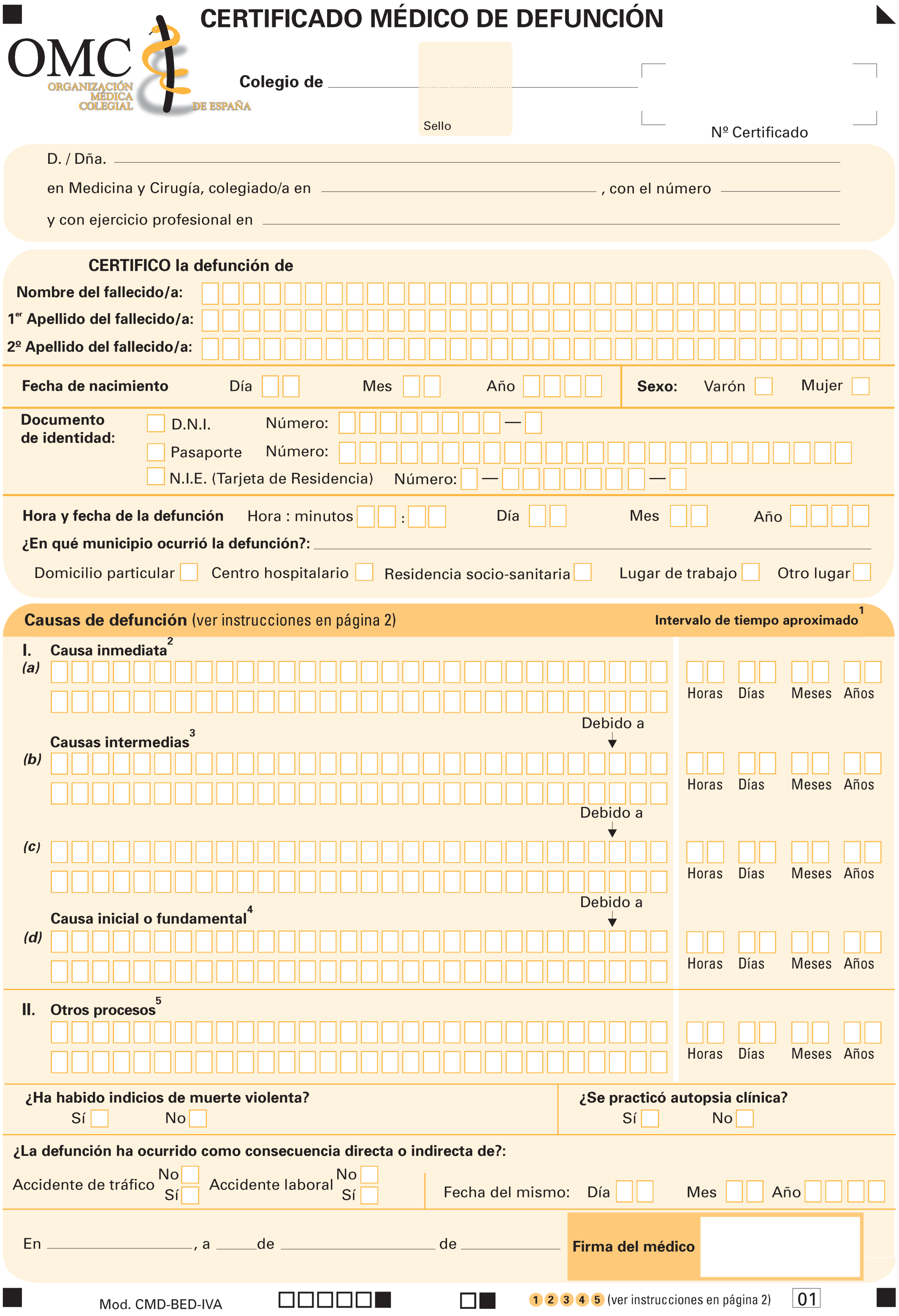
Firma del médico / ***Signatura del metge***

1 2 3 4 5 (ver instrucciones al dorso) / *(vegeu instruccions al dors)* □□□□□■

Mod. CMD-BED-C

01

Figure E.10: Certificado médico de defunción (version 2009)



CERTIFICADO MÉDICO DE DEFUNCIÓN

OMC ORGANIZACIÓN MÉDICA COLEGIAL DE ESPAÑA

Colegio de ____________________

Sello

Nº Certificado

CLASE 3ª SERIE A

3,54 Euros. Derechos autorizados, I.V.A. incluido

D. / Dña. ____________________

en Medicina y Cirugía, colegiado/a en ____________________ , con el número ____________

y con ejercicio profesional en ____________________

CERTIFICO la defunción de

Nombre del fallecido/a: □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

1er Apellido del fallecido/a: □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

2º Apellido del fallecido/a: □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

Fecha de nacimiento Día □□ Mes □□ Año □□□□ **Sexo:** Varón □ Mujer □

Documento de identidad:
□ D.N.I. Número: □□□□□□□□ — □
□ Pasaporte Número: □□□□□□□□□□□□□□□□□□□□□□□□□
□ N.I.E. (Tarjeta de Residencia) Número: □ — □□□□□□□ — □

Hora y fecha de la defunción Hora : minutos □□ : □□ Día □□ Mes □□ Año □□□□

¿En qué municipio ocurrió la defunción?: ____________________

Domicilio particular □ Centro hospitalario □ Residencia socio-sanitaria □ Lugar de trabajo □ Otro lugar □

Causas de defunción (ver instrucciones en página 2) **Intervalo de tiempo aproximado**[1]

I. Causa inmediata[2]

(a) □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□ □□ Horas □□ Días □□ Meses □□ Años
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

Debido a ↓

Causas intermedias[3]

(b) □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□ □□ Horas □□ Días □□ Meses □□ Años
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

Debido a ↓

(c) □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□ □□ Horas □□ Días □□ Meses □□ Años
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

Debido a ↓

Causa inicial o fundamental[4]

(d) □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□ □□ Horas □□ Días □□ Meses □□ Años
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

II. Otros procesos[5]

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□ □□ Horas □□ Días □□ Meses □□ Años
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

¿Ha habido indicios de muerte violenta? Sí □ No □

¿Se practicó autopsia clínica? Sí □ No □

¿La defunción ha ocurrido como consecuencia directa o indirecta de?:

Accidente de tráfico No □ Sí □ Accidente laboral No □ Sí □ Fecha del mismo: Día □□ Mes □□ Año □□□□

En ____________________, a ______de ____________________ de ____________

Firma del médico

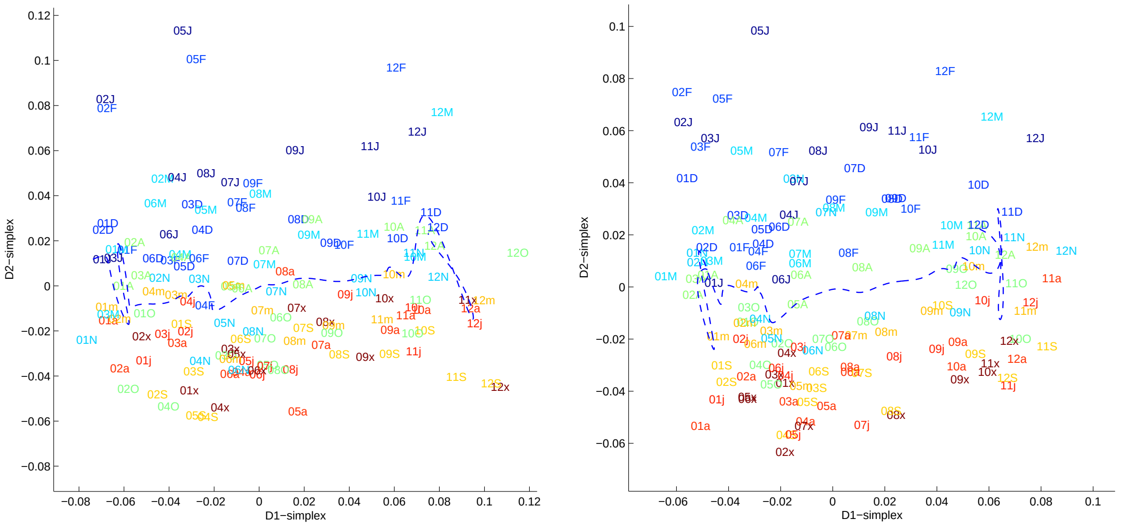
Mod. CMD-BED-IVA □□□□□■ □■ 1 2 3 4 5 (ver instrucciones en página 2) 01

Figure E.11: 3. Certificado médico de defunción (version 2009 - Fixed)

E.8 Temporal variability of basic cause of death

This section shows the IGT plots and temporal heat maps of the basic cause of death for males and females (*BasicCause* variable). The basic cause is the main indicator of causes-of-death in Public Health studies. As a consequence, its quality is generally controlled, and the abrupt changes due to the certificate change in 2009 were retrospectively corrected. However, slight differences can still be found due to this change: first, in the IGT plots as a slight break in the temporal flow in month 99, and second, in the temporal heat maps as slight abrupt changes in the frequencies of some causes in 2009.

Additionally, it is remarkable the strong seasonal effect in the basic cause of death, observed in the IGT plots as the component associated to the color temperature (cold and warm colors for winter and summer periods, respectively), and in the absolute frequencies temporal heat maps as the periodic peaks of frequencies in causes-of-death. This is mainly due to the seasonality of diseases, mainly winter-specific respiratory diseases and summer heart diseases.



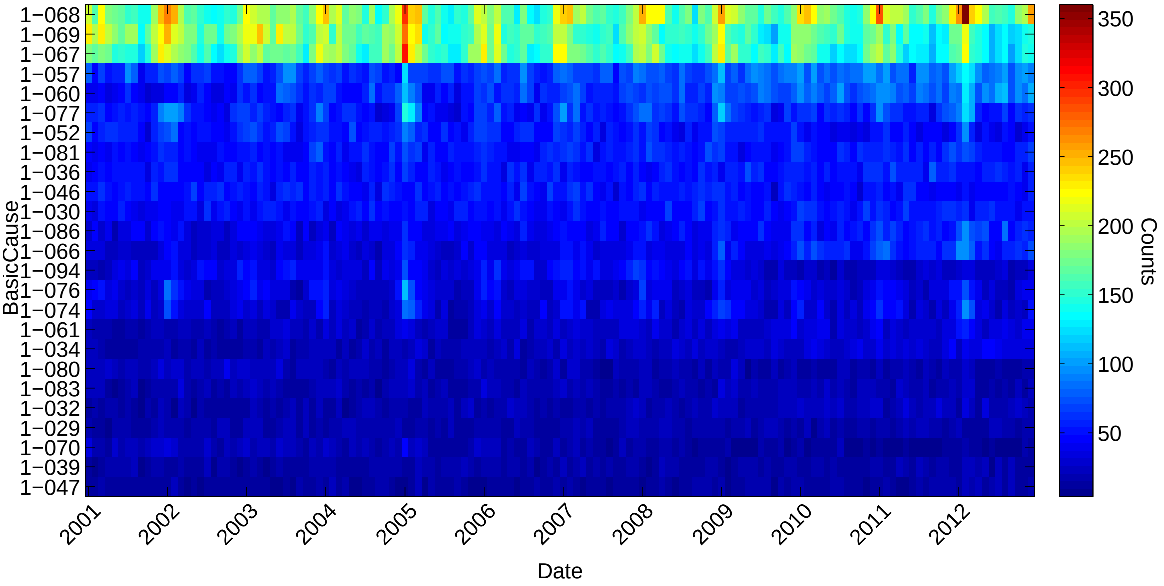
(a) Females

(b) Males

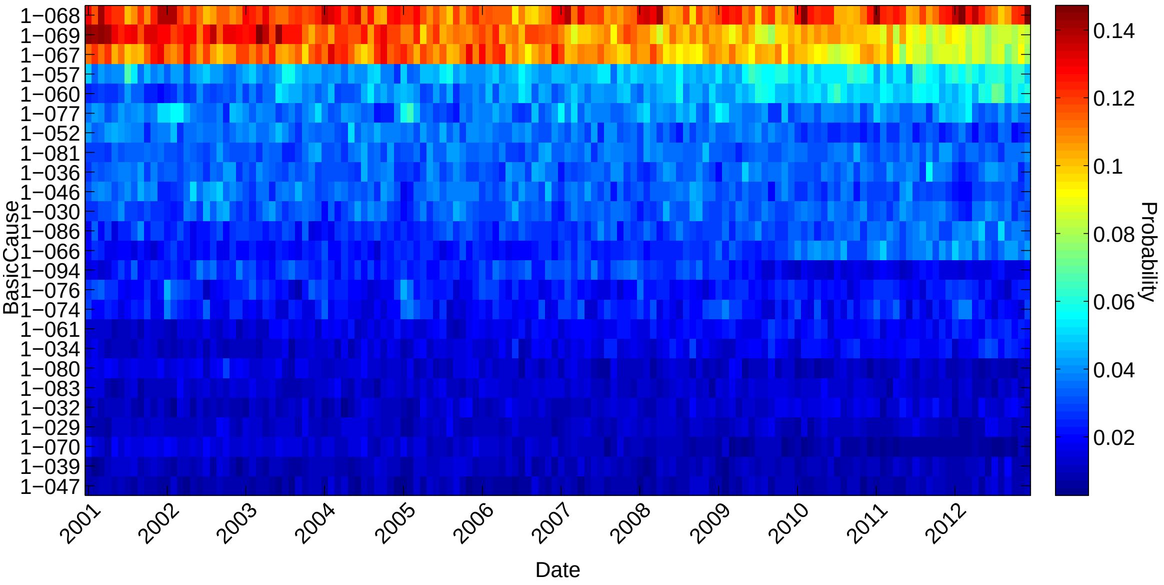
Figure E.12: IGT plots of basic cause of death for both sexes

E.9 Temporal heatmaps of age at death

The figures in this section show the absolute and PDF temporal heat maps of the Age of death for females (Figure E.15) and males (Figure E.16). The most remarkable finding is a gradual change related to an increase in life expectancy. Additionally, detailed effects in specific age ranges can be observed, such as a linear evolution of a population gap coinciding with the lack of newborns during the Spanish War in 1938 or the increase of deaths in 2002 and 2005 due to flu.

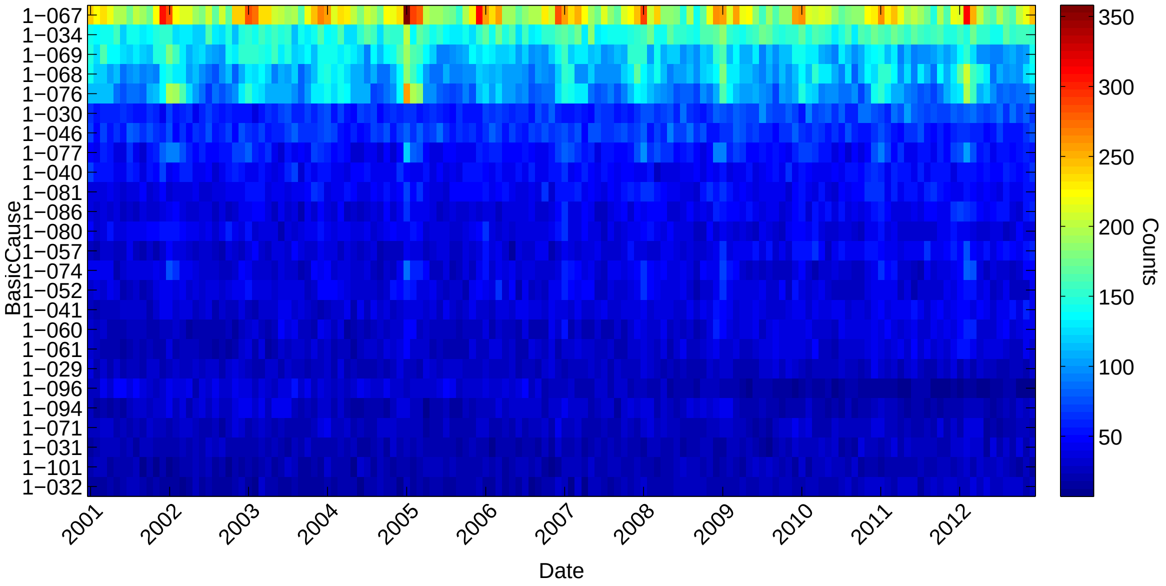


(a) Absolute frequencies temporal heat map

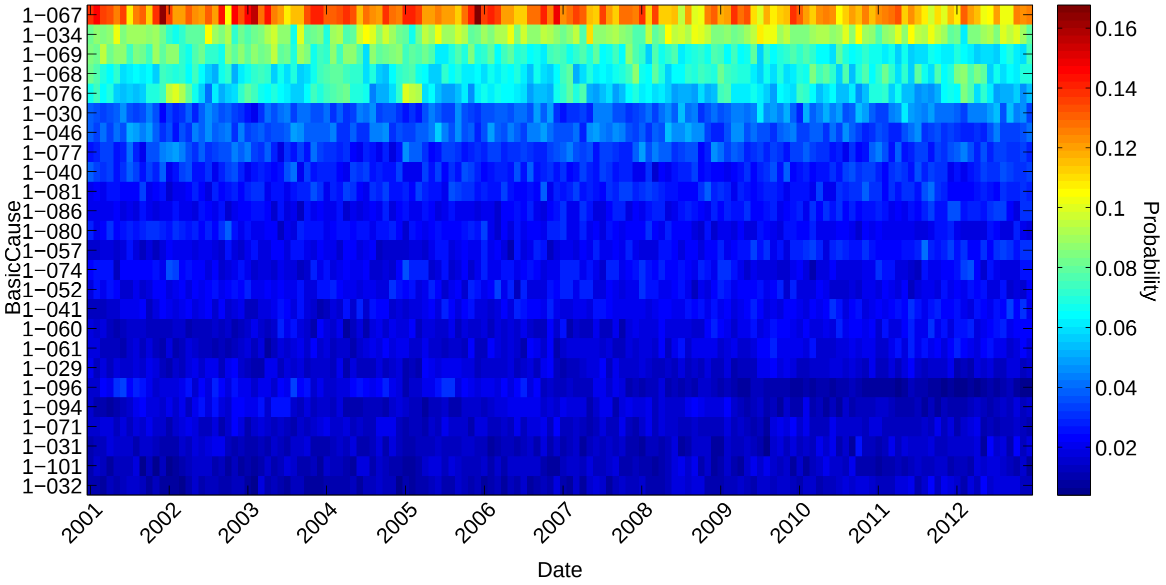


(b) PDF temporal heat map

Figure E.13: Temporal heat maps of basic cause of death for females (see codes in Table E.1)

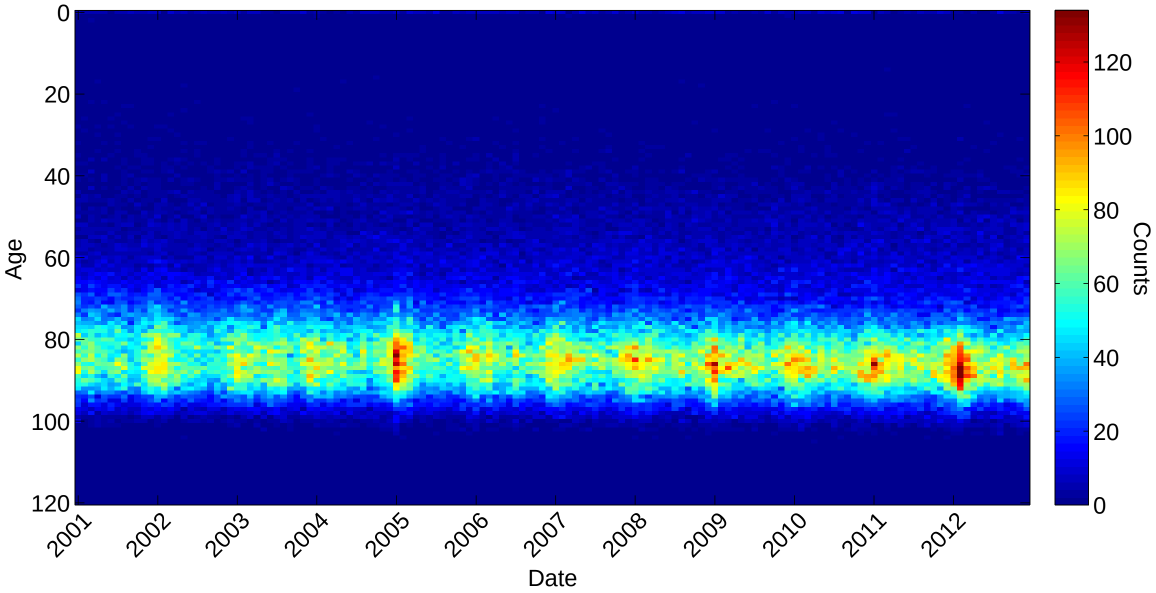


(a) Absolute frequencies temporal heat map

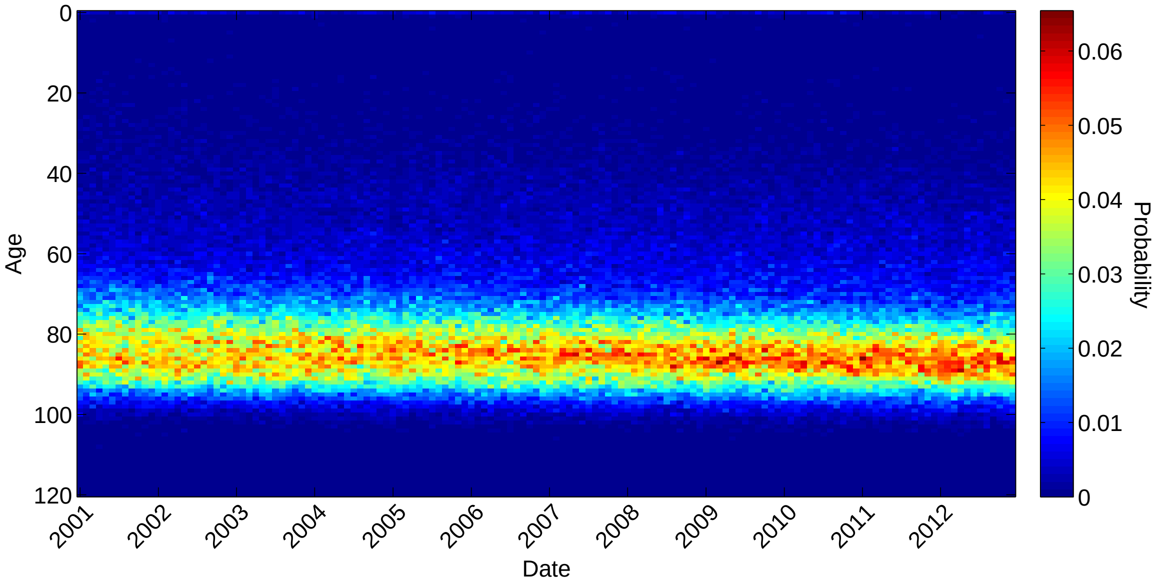


(b) PDF temporal heat map

Figure E.14: Temporal heat maps of basic cause of death for males (see codes in Table E.1)

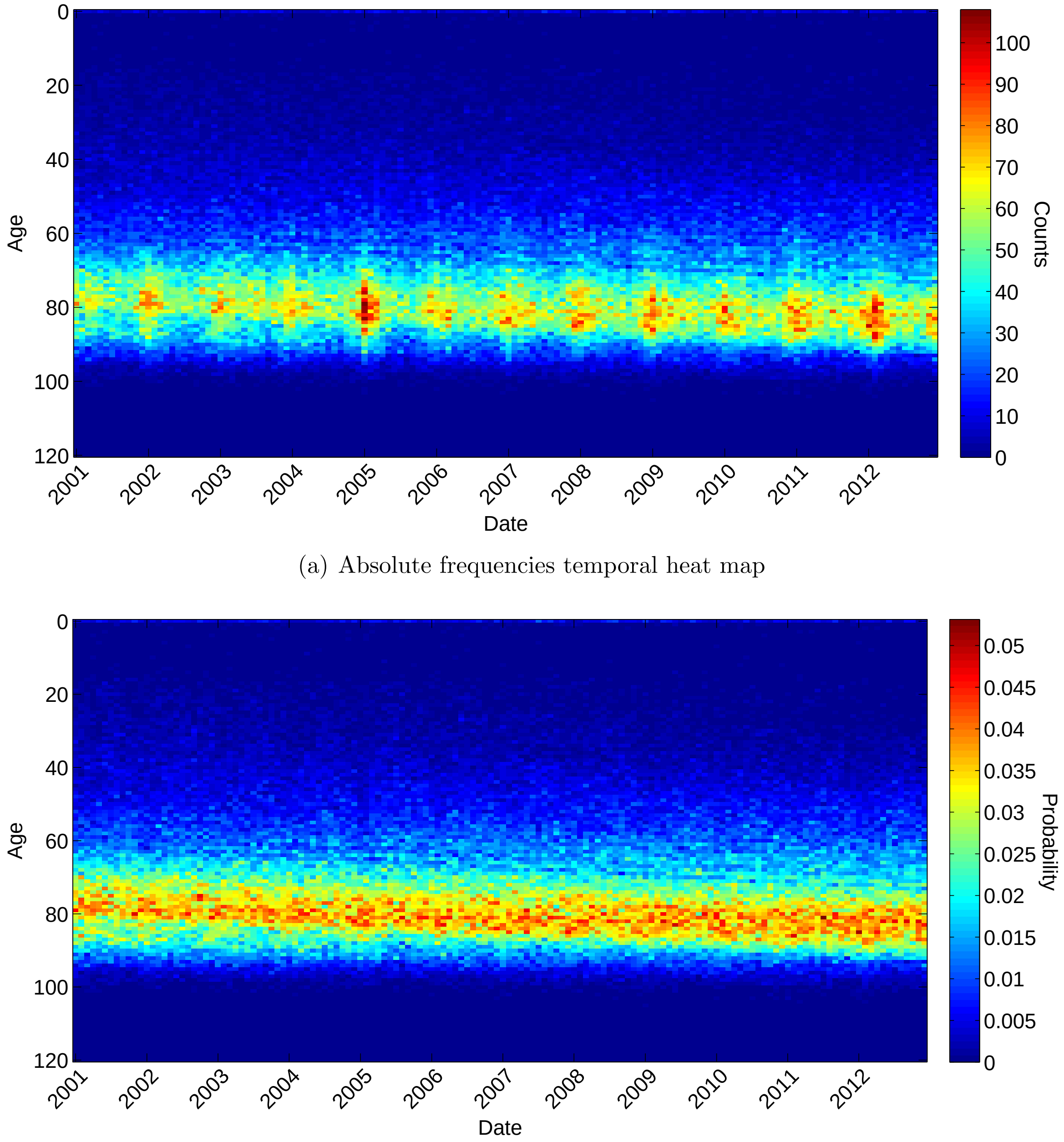


(a) Absolute frequencies temporal heat map



(b) PDF temporal heat map

Figure E.15: Temporal heat maps of age at death for females



(a) Absolute frequencies temporal heat map

(b) PDF temporal heat map

Figure E.16: Temporal heat maps of age at death for males