Métodos espacio-temporales probabilísticos para el control de calidad de datos biomédicos, aplicación al Registro de Mortalidad de la Comunitat Valenciana

Programa de Doctorado en Tecnologías para la Salud y el Bienestar

Doctorando: Carlos Sáez Silvestre

Directores: Montserrat Robles Viejo, Juan Miguel García Gómez

Grupo de Informática Biomédica (IBIME), Instituto de Tecnologías de la Información y Comunicaciones ITACA, Universitat Politècnica de València

Junio de 2015

Los procesos de toma de decisiones ejecutivas, estratégicas y resultados de estudios de investigación dependen actualmente de la información almacenada en sistemas computacionales. La falta de calidad en dichos datos es un problema particularmente importante en la información biomédica, que puede tener consecuencias directas o indirectas en la salud pública, atención sanitaria de los pacientes, u obstaculizar la reutilización de datos para investigación, ensayos clínicos o políticas sanitarias. Esto es debido principalmente a que los Sistemas de Información Clínica generalmente ofrecen niveles de calidad de datos inapropiados para su uso secundario. La mayoría de estudios sobre calidad de datos se han centrado en análisis semánticos o estándares de información clínica. Sin embargo, la variabilidad en los datos que puede existir entre diferentes fuentes de datos (hospitales, profesionales, etc.) o a lo largo del tiempo ha recibido poca atención. Estos problemas comienzan a ser importantes en la era del *Big Data*, donde crecen los grandes repositorios de datos multicéntricos. Además, los métodos estadísticos clásicos pueden no ser adecuados en entornos *Big Data* biomédicos más aun considerando datos multivariantes, multimodales y con múltiples tipos de variables.

El objetivo de esta tesis es profundizar en el diseño y evaluación de métodos y métricas para el control de la variabilidad entre fuentes (espacial) y en el tiempo (temporal) constituyendo una metodología sistemática para el control de calidad de datos. Para ser robustos a los problemas anteriores, los métodos se basan en un marco de Teoría y Geometría de Información, basado en la inferencia de variedades de Riemann estadísticas no paramétricas a partir de distancias normalizadas entre distribuciones de probabilidad, ya sea entre diferentes fuentes de datos, a lo largo del tiempo o de forma espacio-temporal. La variedad espacial de dimensión completa da lugar a un simplex geométrico a partir del cual se obtienen métricas de desviación probabilística global (GPD) y anomalía de fuente (SPO). La variedad temporal permite proyectar la evolución de conceptos probabilísticos en el tiempo, y construir un Control Estadístico de Procesos probabilístico (PDF-SPC) que monitoriza la distribución Beta de las distancias acumuladas. Finalmente, la variedad espacio-temporal permite monitorizar la evolución de agrupaciones y anomalías multifuente. Cada método ha sido evaluado con registros públicos como el registro de altas hospitalarias de EEUU.

Con el fin de validar el uso sistemático de los nuevos métodos en conjunto éstos se han aplicado para analizar y evaluar la calidad de datos del Registro de Mortalidad de la Comunitat Valenciana. Se han encontrado hallazgos como la partición temporal del registro en dos

subgrupos temporales asociados a un cambio en el Certificado de Defunción, anomalías mensuales debidas a datos incompletos, grupos de Departamentos de Salud con prácticas de codificación aisladas, Departamentos anómalos y el efecto de reasignación de Departamentos.

Actualmente la tesis ha dado lugar a tres artículos en revistas de alto impacto en estadística e informática y dos en congresos internacionales relevantes en informática médica. Un último artículo se encuentra actualmente en revisión.

Finalmente, los resultados intermedios de la tesis han permitido obtener un proyecto Retos-Colaboración del MINECO para su industrialización, llevada a cabo de forma conjunta entre la UPV y la Spin-off Veratech for Health.

Métodos espacio-temporales probabilísticos para el control de calidad de datos biomédicos, aplicación al Registro de Mortalidad de la Comunitat Valenciana

Programa de Doctorado en Tecnologías para la Salud y el Bienestar

Carlos Sáez (doctorando), Montserrat Robles, Juan M García-Gómez (directores)

Grupo de Informática Biomédica (IBIME), Instituto de Tecnologías de la Información y Comunicaciones ITACA
Universitat Politècnica de València

1. Objetivos

Objetivo general

Mejorar eficacia y eficiencia investigación biomédica y toma de decisiones sanitarias mediante una óptima reutilización datos con **información válida y fiable**

Objetivos específicos

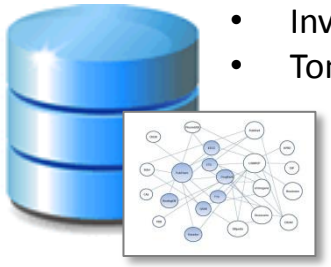
- Controlar variabilidad de datos entre fuentes y en el tiempo (espacio-temporal)
- Diseñar métodos robustos a **Big Data** y datos multi-variantes, multi-modales y multi-tipo

2. Antecedentes

Sistemas de Información Clínica

Reutilización datos

Sistemas de conocimiento



- Investigación
- Toma de decisiones

*"Un 70-90% de datos con **problemas** en repositorios investigación"*

*"La preparación y limpieza datos supone el 80% del **coste** de proyectos de reutilización"*

"Variabilidad entre fuentes (p.ej., hospitales y profesionales) y a lo largo del tiempo"

- **Análisis ineficientes**
- **Modelos o hipótesis sesgados**
- **Decisiones subóptimas**

3. Métodos desarrollados

Marco probabilístico no-paramétrico:

$$\mathcal{M} \leftarrow Embedding(\mathbf{D} : D_{ij'} = JS(p_i||p'_j)^{-1/2})$$

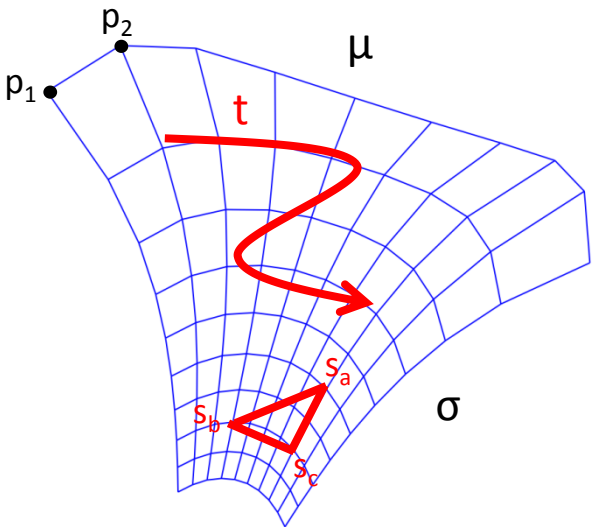
Variabilidad espacial

- *Global Probabilistic Deviation* (***GPD***)
- *Source Probabilistic Outlyingness* (***SPO***)
- Proyección simplicial disimilaridad

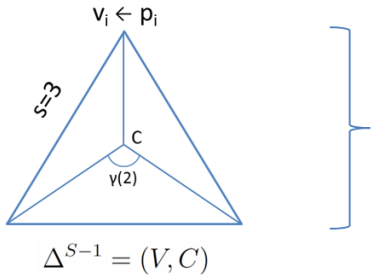
Variabilidad temporal

- Proyección variedad de Riemann temporal (***IGT-plot***)
- *Probabilistic Statistical Process Control* (***PDF-SPC***)
- Monitorización métricas espacio-temporal

Ejemplo p~N(μ, σ)



$$\mathcal{M}_{\mu,\sigma} = \{p(x|\mu,\sigma)\}$$



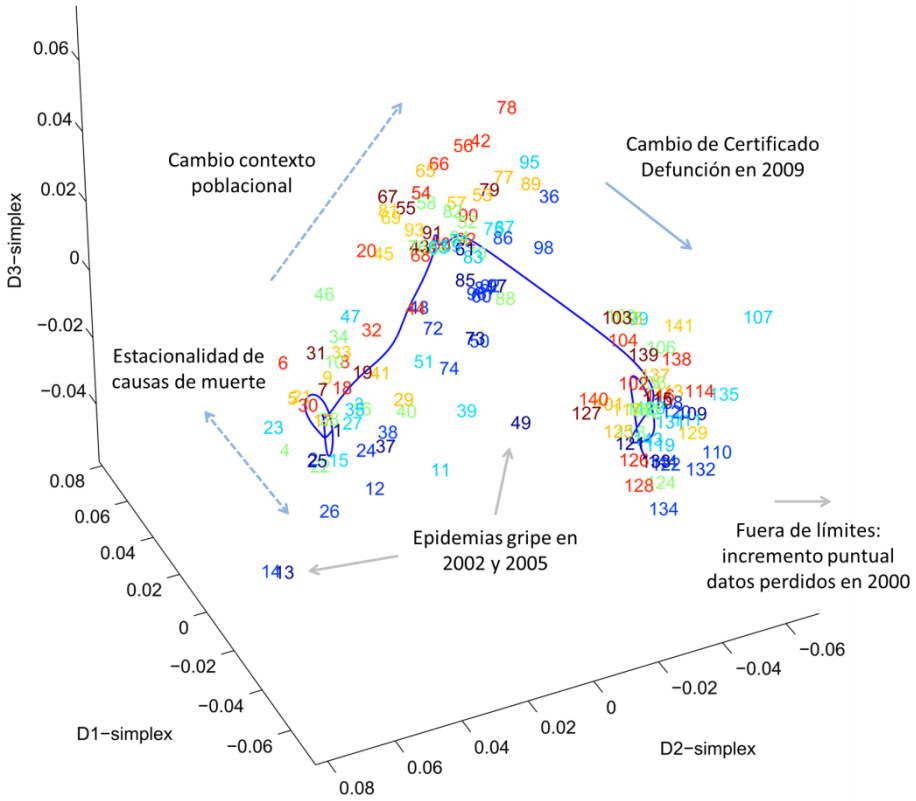
$$\Delta^{S-1} = (V, C)$$

$$GPD = \frac{2sin(\gamma(D)/2)\sum_{s=1}^{S} d(V_s, C)}{S}$$

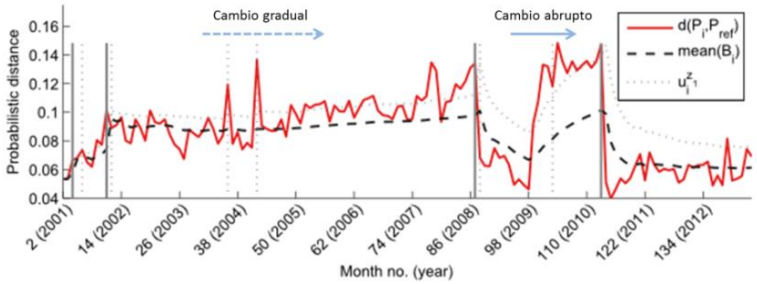
$$SPO_s = \frac{d(V_s, C)}{1 - 1/S}$$

4. Resultados en Registro de Mortalidad de la Comunitat Valenciana (1/3)

Variabilidad temporal



Evolución probabilística Registro de Mortalidad multivariante completo mediante variedad de Riemann estadística temporal (***IGT-plot***)



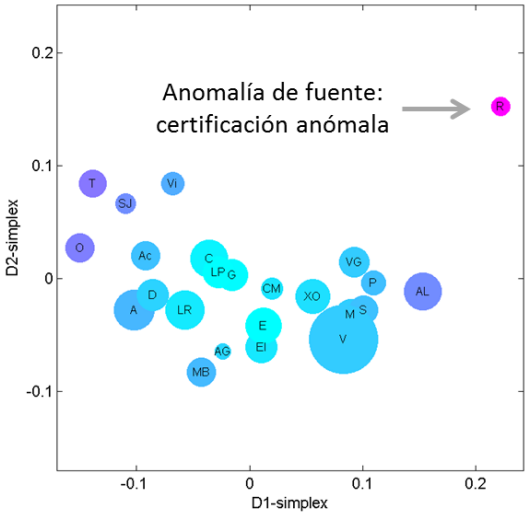
Probabilistic Statistical Process Control (***PDF-SPC***)

Modelado estadístico o hipótesis incompatibles entre antes y después del cambio de Certificado de Defunción

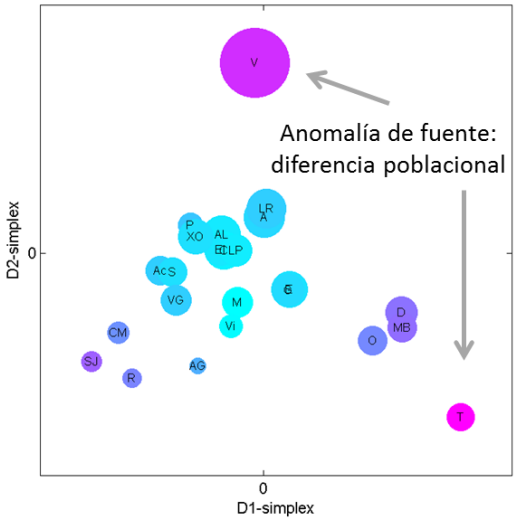
*N = 512143 defunciones entre 2000-2012 en 24 Departamentos de Salud

4. Resultados en Registro de Mortalidad de la Comunitat Valenciana (2/3)

Variabilidad espacial



Multicausas certificado defunción



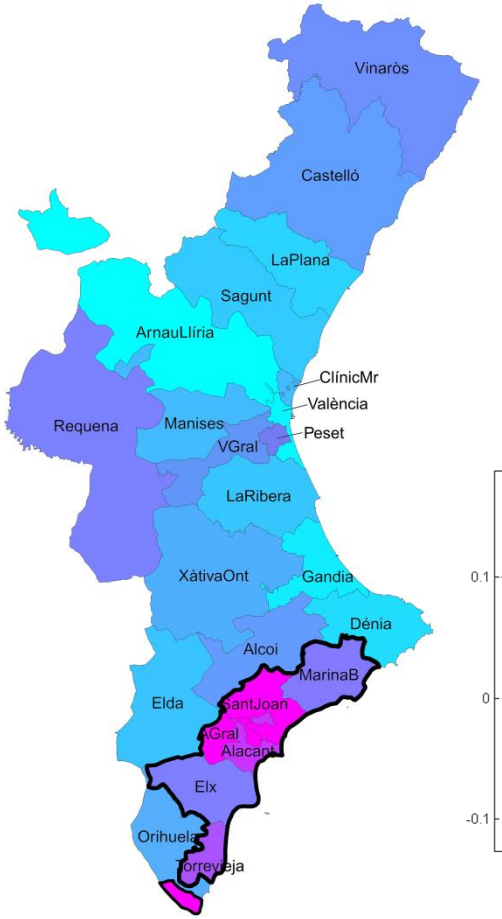
Edad+Sexo+Causa básica de muerte

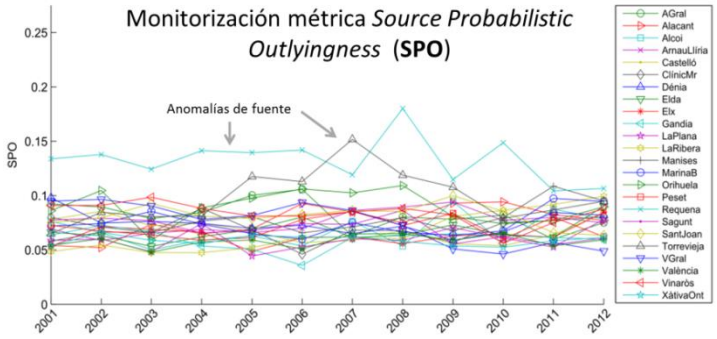
2D-simplices disimilaridad espacial multivariante en 2008

Modelado estadístico o decisiones de salud incompatibles entre Departamentos

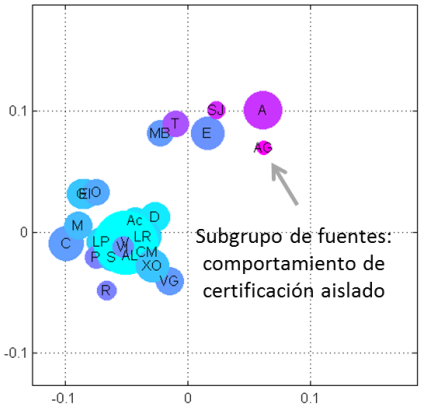
4. Resultados en Registro de Mortalidad de la Comunitat Valenciana (3/3)

Variabilidad espacio-temporal

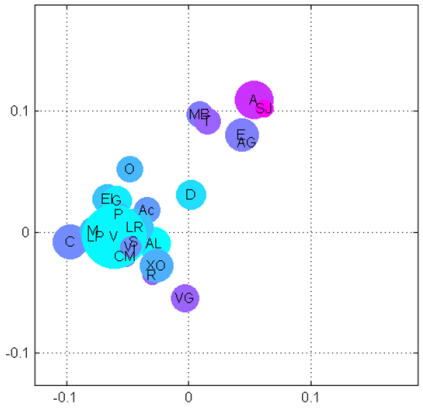




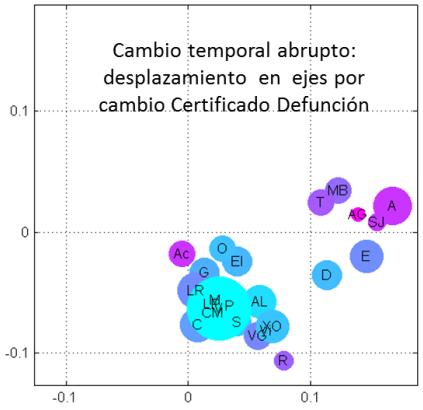
2001-2004



2005-2008



2009-2012



2D-simplices espacio-temporales Causa Intermedia de muerte 2001-2012

5. Utilidad

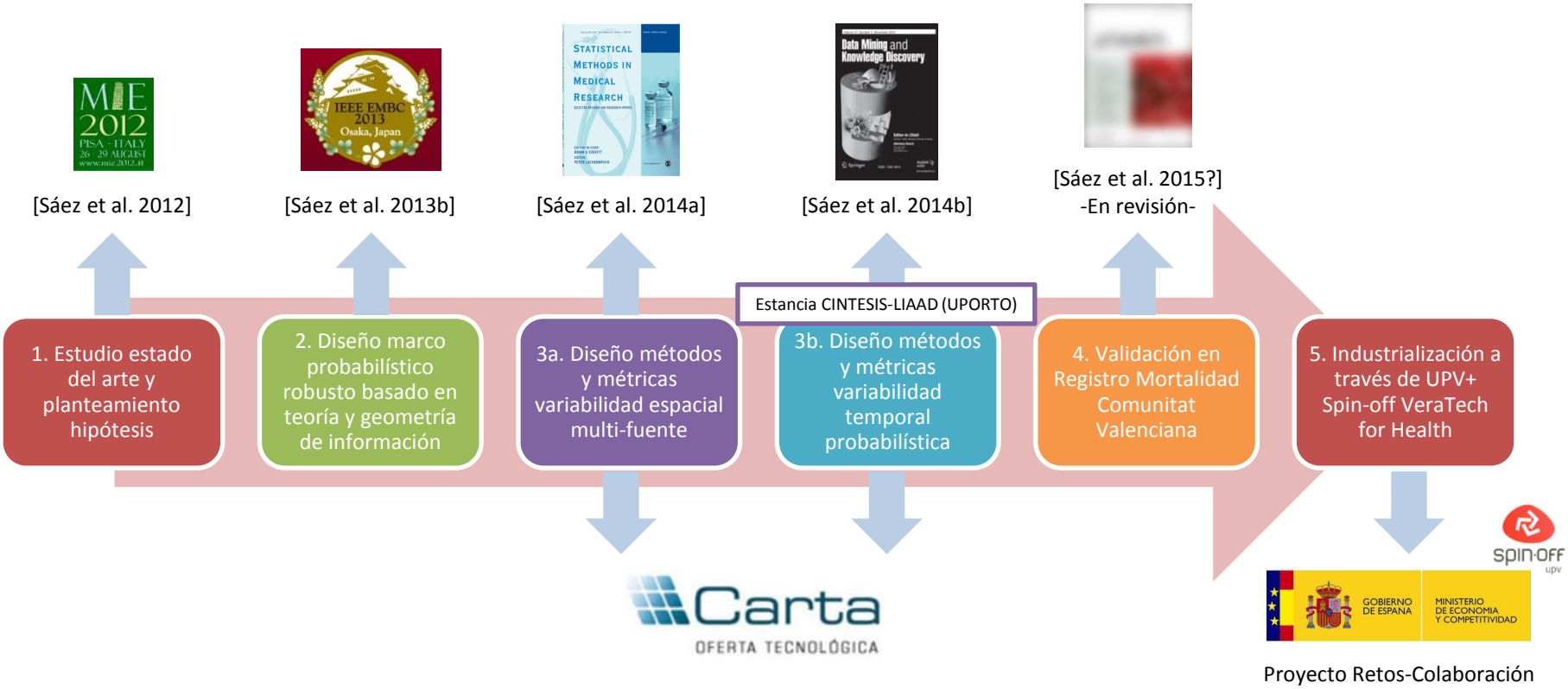
Usos principales:

- Facilitar el uso de ***Big Data*** como ***Right Data***
- Cuantificar la calidad de repositorios de investigación multi-céntricos
- Monitorizar servicios sanitarios
- Controlar variabilidad en ensayos clínicos
- Certificación y rating de calidad de datos

Usuarios potenciales

- Instituciones científicas
- Proveedores de salud
- Gestores sanitarios
- Industria farmacéutica
- Certificadoras calidad

6. Etapas y desarrollo investigación



[Sáez et al. 2012] **C Sáez**, J Martínez-Miranda, M Robles & JM García-Gómez. Organizing data quality assessment of shifting biomedical data. Studies in Health Technology and Informatics, 180, 721–725.
[Sáez et al. 2013b] **C Sáez**, M Robles, JM García-Gómez. Comparative study of probability distribution distances to define a metric for the stability of multi-source biomedical research data. Conference Proceedings IEEE Engineering in Medicine and Biology Society. 2013:3226-9. 2013.
[Sáez et al. 2014a] **C Sáez**, M Robles & JM 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 Press).
[Sáez et al. 2014b] **C Sáez**, M Robles & JM García-Gómez. Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. Data Mining and Knowledge Discovery. Published Online First (In Press).
[Sáez et al. 2015?] **C Sáez**, et al. Applying probabilistic temporal and multi-site data quality control methods for reuse data: findings in a public health mortality registry in a region of Spain and systematic approach. Under review.

Carlos Sáez

Grupo de Informática Biomédica (IBIME)
Instituto de Tecnologías de la Información y Comunicaciones ITACA
Universitat Politècnica de València

carsaesi@ibime.upv.es

Gracias por su atención