Alvaro J. Riascos Villegas
University of los Andes y Quantil

Octubre 30, 2017

Contents

Introducción

- Presento los ejemplos más significativos de aplicaciones exitosas de las Matemáticas Aplicadas en la acedemía e industría: **una visión muy personal** como profesor de UNIANDES y co-Director de Quantil.
- Estas son aplicaciones a:
  1. Detección de minería ilegal.
  2. Minería de datos médicos.
  3. Mercadeo basado en datos.
  4. Predicción del crimen.
  5. Mineria de texto.
  6. Políticas públicas.

Introducción

- Presento los ejemplos más significativos de aplicaciones exitosas de las Matemáticas Aplicadas en la acedemía e industría: **una visión muy personal** como profesor de UNIANDES y co-Director de Quantil.
- Estas son aplicaciones a:
  1. Detección de minería ilegal.
  2. Minería de datos médicos.
  3. Mercadeo basado en datos.
  4. Predicción del crimen.
  5. Mineria de texto.
  6. Políticas públicas.

Introducción

- Presento los ejemplos más significativos de aplicaciones exitosas de las Matemáticas Aplicadas en la acedemía e industría: **una visión muy personal** como profesor de UNIANDES y co-Director de Quantil.
- Estas son aplicaciones a:
  1. Detección de minería ilegal.
  2. Minería de datos médicos.
  3. Mercadeo basado en datos.
  4. Predicción del crimen.
  5. Mineria de texto.
  6. Políticas públicas.

Introducción

- Presento los ejemplos más significativos de aplicaciones exitosas de las Matemáticas Aplicadas en la acedemía e industría: **una visión muy personal** como profesor de UNIANDES y co-Director de Quantil.
- Estas son aplicaciones a:
  1. Detección de minería ilegal.
  2. Minería de datos médicos.
  3. Mercadeo basado en datos.
  4. Predicción del crimen.
  5. Mineria de texto.
  6. Políticas públicas.

Introducción

- Presento los ejemplos más significativos de aplicaciones exitosas de las Matemáticas Aplicadas en la acedemía e industría: **una visión muy personal** como profesor de UNIANDES y co-Director de Quantil.
- Estas son aplicaciones a:
  1. Detección de minería ilegal.
  2. Minería de datos médicos.
  3. Mercadeo basado en datos.
  4. Predicción del crimen.
  5. Mineria de texto.
  6. Políticas públicas.

Introducción

- Presento los ejemplos más significativos de aplicaciones exitosas de las Matemáticas Aplicadas en la acedemía e industría: **una visión muy personal** como profesor de UNIANDES y co-Director de Quantil.
- Estas son aplicaciones a:
  1. Detección de minería ilegal.
  2. Minería de datos médicos.
  3. Mercadeo basado en datos.
  4. Predicción del crimen.
  5. Mineria de texto.
  6. Políticas públicas.

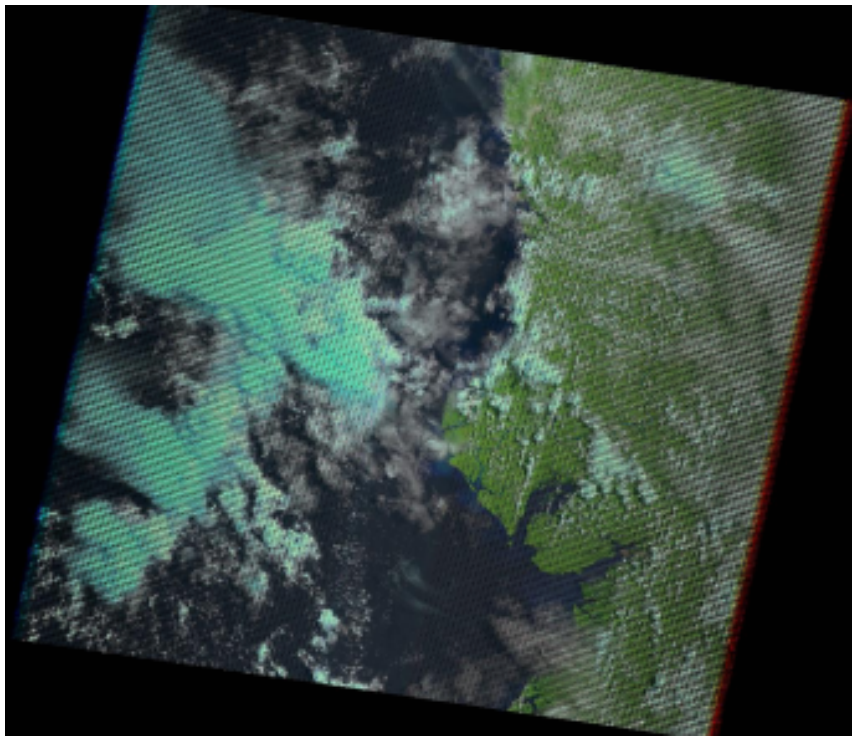
Introducción

- Presento los ejemplos más significativos de aplicaciones exitosas de las Matemáticas Aplicadas en la acedemía e industría: **una visión muy personal** como profesor de UNIANDES y co-Director de Quantil.
- Estas son aplicaciones a:
  1. Detección de minería ilegal.
  2. Minería de datos médicos.
  3. Mercadeo basado en datos.
  4. Predicción del crimen.
  5. Mineria de texto.
  6. Políticas públicas.

Contents

Detección de minería ilegal usando imágenes satelitales

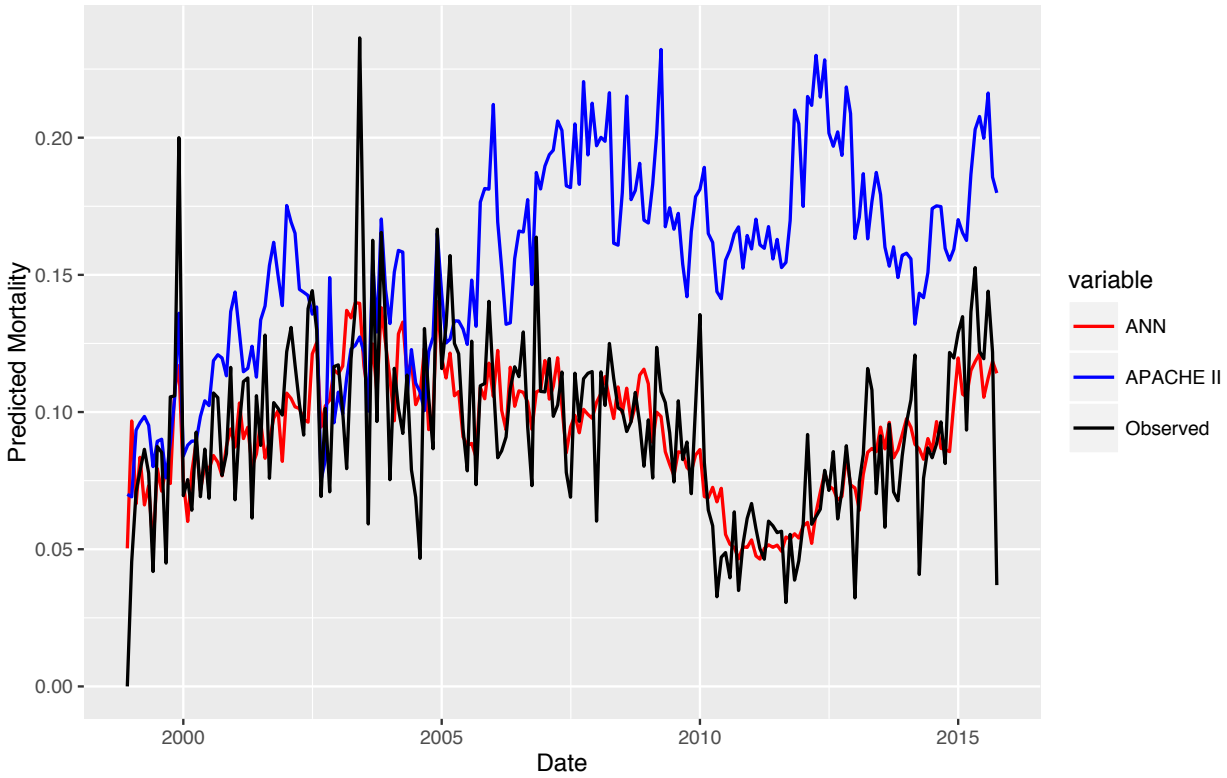
- Cortesía de Santiago Saavedra (Universidad del Rosario)



Contents

Minería de datos médicos

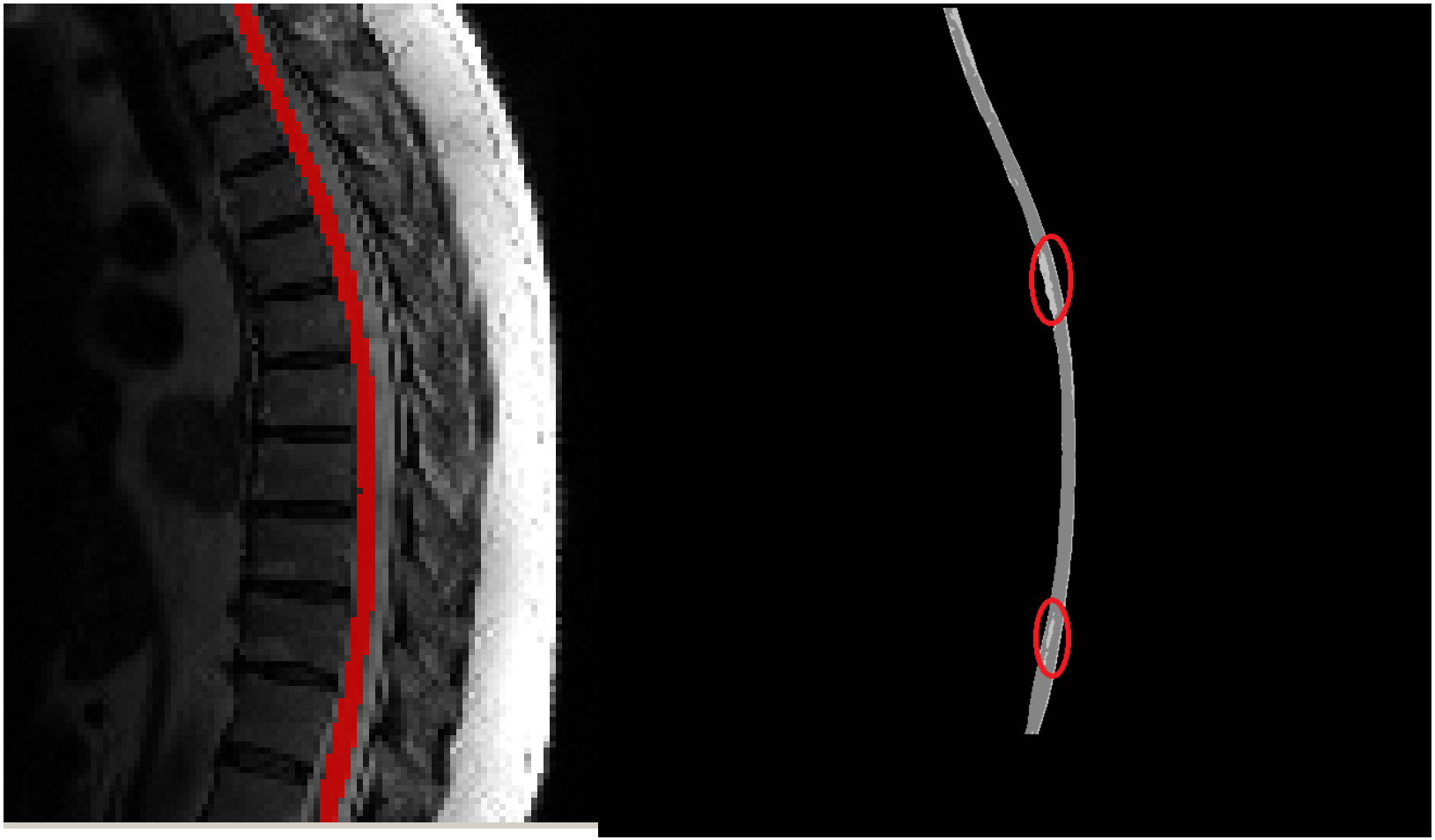
- Cortesía Fundación Valle del Lili (FVL).



Imágenes

Gráfica: Cortesía de Johns Hopkins University

(a) Spinal Cord (b) Spinal Cord Injury

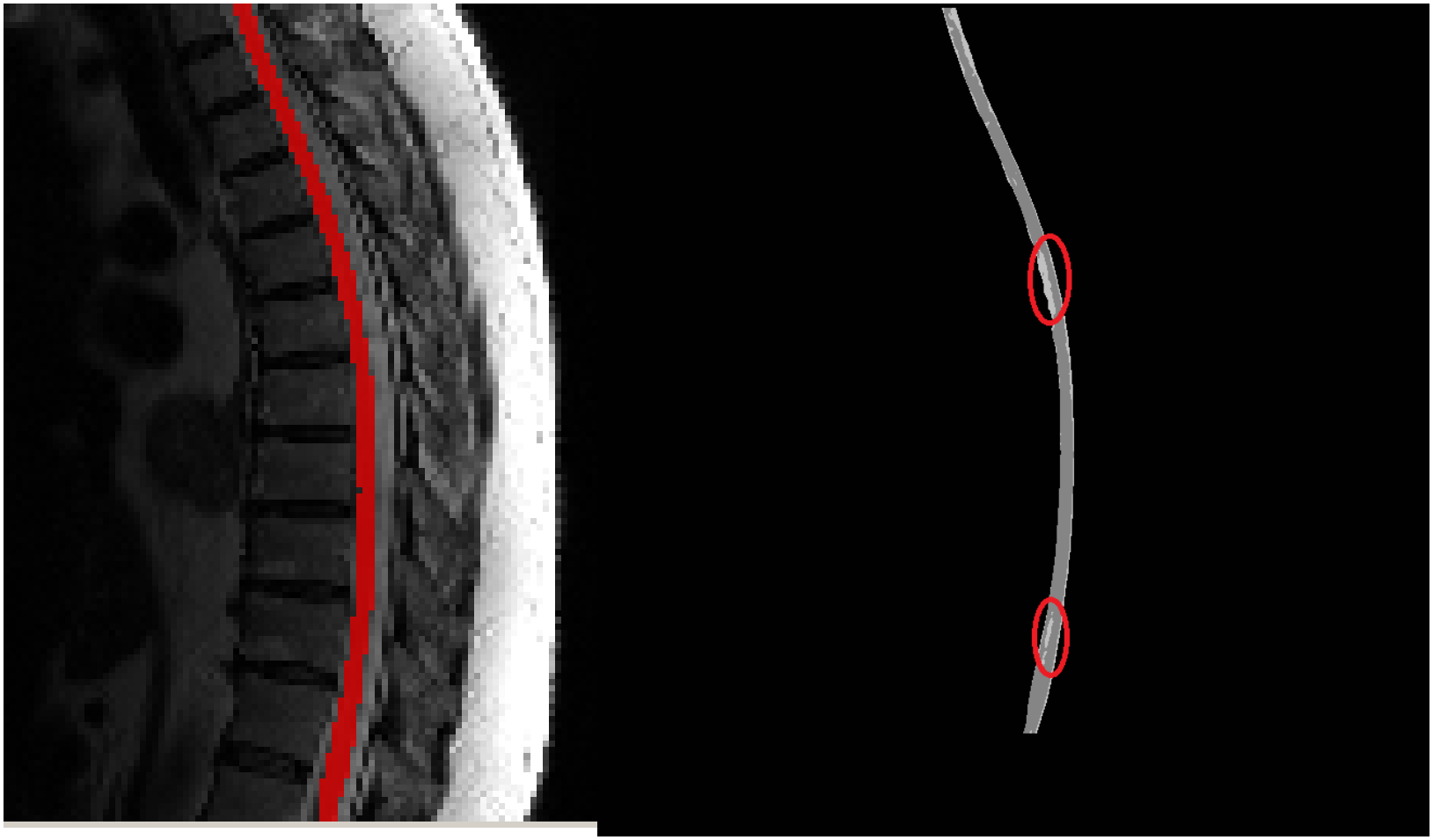


- Detección de signos vitales sin contacto: Uniandes, Harvard, ICESI, EVI

Imágenes

Gráfica: Cortesía de Johns Hopkins University

(a) Spinal Cord (b) Spinal Cord Injury

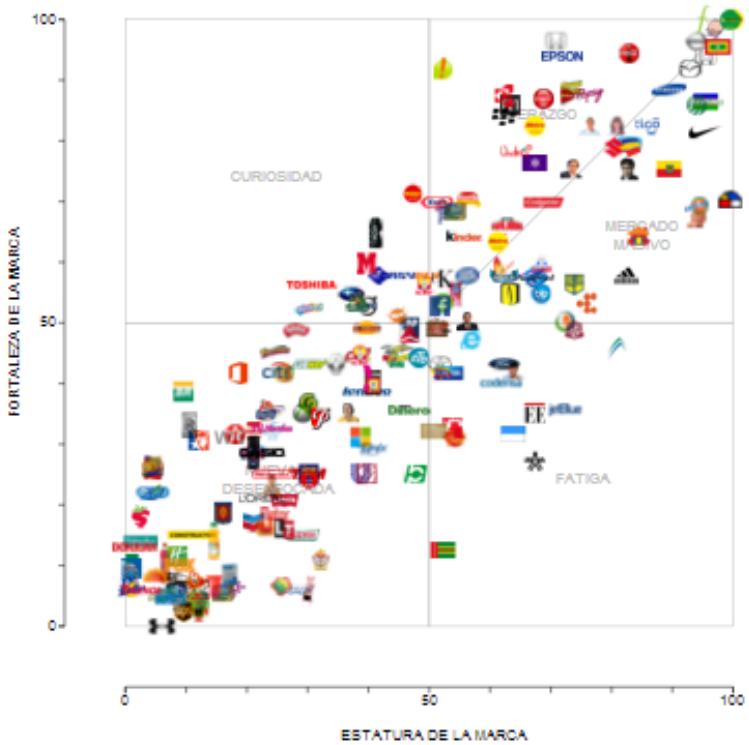


- Detección de signos vitales sin contacto: Uniandes, Harvard, ICESI, EVI

Contents

Mercadeo basado en datos

- Cortesía: Barbara & Frick y Young & Rubican.
- Minería de redes sociales (twitter).



Contents

Modelo Espacio - Temporal

- Este es un modelo basado en la clasificación de los eventos como antecedentes y réplicas.
- Es el estado del arte en modelos de predicción del crimen.
- Está motivado por teorías de contagio espacial y réplicas temporales.

Modelo Espacio - Temporal: Motivación

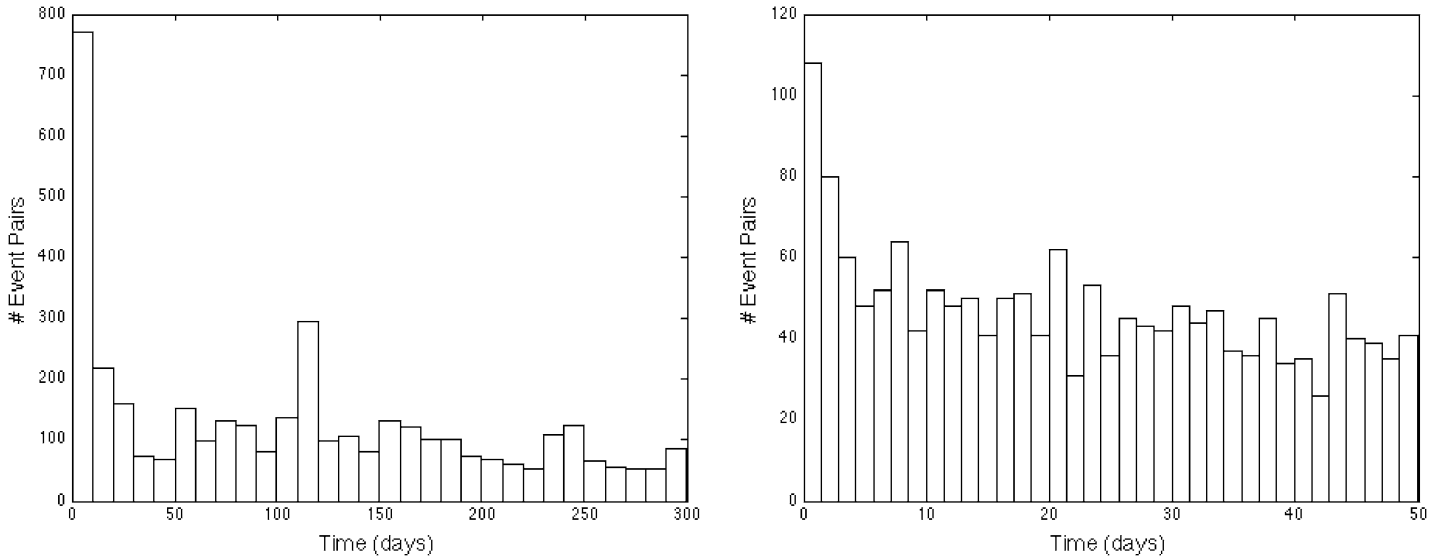


Figure 1. On the left, histogram of times (less than 300 days) between Southern California earthquake events of magnitude 3.0 or greater separated by 110 kilometers or less. On the right, histogram of times (less than 50 days) between burglary events separated by 200 meters or less.

Modelo Espacio - Temporal: Motivación

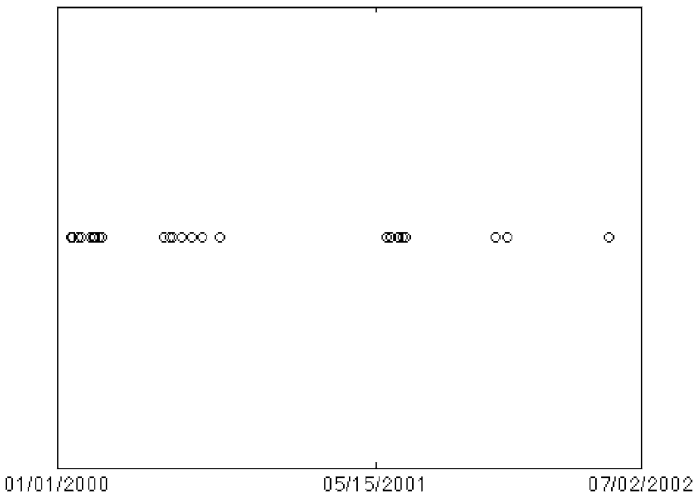


Figure 2. Times of violent crimes between two rivalry gangs in Los Angeles.

Modelo Espacio - Temporal: especificación

- Se considera un modelo de la intensidad espacio temporal del crimen de la forma:

$$\lambda(t,x,y) = \mu(t,x,y) + \sum_{k:t_k<t} g(t-t_k, x-x_k, y-y_k) \quad (1)$$

- Sea $\mu(t,x,y) = \nu(t)\mu(x,y)$

Modelo Espacio - Temporal: especificación

- Suponiendo que el modelo es correcto la probablidad de que un evento i sea un evento antecedente es:

$$\pi_{ii} = \frac{\nu(t_i)\mu(x_i, y_i)}{\lambda(t_i, x_i, y_i)} \tag{2}$$

- La probabilidad de que el evento j cause el evento i (evento réplica) es:

$$\pi_{ji} = \frac{g(t_i - t_j, x_i - x_j, y_i - y_j)}{\lambda(t_i, x_i, y_i)} \tag{3}$$

Modelo Espacio - Temporal: estimación

- Sea P la matriz de probabilidades (p_{ji}) (obsérvese que la suma de columnas da 1).
- Elija una probabilidad inicial P_0.
- Muestrear de P_0 puntos de transfondo y réplica: $(t_k, x_k, y_k, p_{kk})_{k=1,\ldots N}$ y $(t_i - t_j, x_i - x_j, y_i - y_j, p_{ji})_{i>j}$ y usando esta muestra hacer una estimación inicial de:

  1. μ_1 de μ.
  2. g_1 de g.
  3. ν_1 de ν

  usando KDE con ancho de banda variable.
- Estimar inicalmente λ_1 usando que las columnas de P suman 1.

Modelo Espacio - Temporal: estimación

- En cada iteración n, ajustamos simulados con P_n los datos $g_n(t_i - t_j, x_i - x_j, y_i - y_j, p_{ji})$ de g. Ahora ajustamos una un kernel del siguiente tipo:

$$g_n(t,x,y) = \frac{1}{N}\sum_{i=1}^{N_0} \frac{1}{\sigma_x\sigma_y\sigma_t(2\pi)^{\frac{3}{2}}D_i^3} \times exp(-\frac{(x-x_i)^2)}{2\sigma_x^2 D_i^2} - \frac{(y-y_i)^2)}{2\sigma_y^2 D_i^2} - \frac{(t-t_i)^2)}{2\sigma_t^2 D_i^2})$$

- Usamos un procedimiento similar para estimar ν y μ. En el primer caso un Kernel Gausssiano univariado y el segundo caso un Kernel Gaussiano bivariado.

Modelo Espacio - Temporal: estimación

- Actualizar P_1.
- Repetir los pasos anteriores hasta que la matriz P no cambie mucho.
- El número de probabilidades a estimar en cada iteración es del orden de N^2.
- Esto es computacionalmente constoso (típicamente N del orden de $1,000$ o $30,000$ - Bogota en un año).

Validación

- Utilizamos el *Precision Accuracy Index*

$$\text{PAI} = \frac{\text{Hit Rate}}{\text{Percentage of Area}}$$

$$\text{Hit Rate} = \frac{\text{Crimes predicted in Hotspots}}{\text{Total Crimes}}$$

$$\text{Percentage Area} = \frac{\text{Area of Hotspots}}{\text{Total Area}}$$

- Sin embargo en modelos muy granulares no es una buena medida.

Modelo Espacio - Temporal: Validación

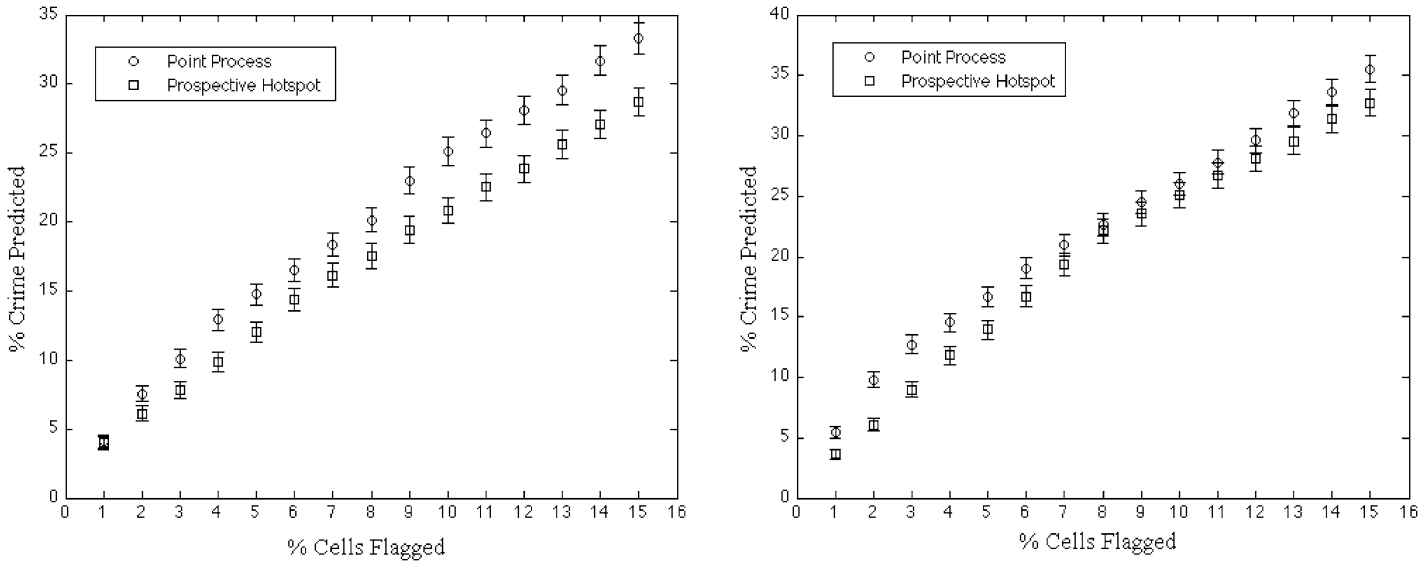


Figure 6. Forecasting strategy comparison. Average daily percentage of crimes predicted plotted against percentage of cells flagged for 2005 burglary using 200 m by 200 m cells. Error bars correspond to the standard error. Prospective hotspot cutoff parameters are 400 meters and 8 weeks (left) and optimal parameters (right) are 200 meters and 39 weeks. Spatial background intensity $\mu(x, y)$ smoothing bandwidth for the point process is 300 meters (left) selected by cross validation and 130 meters (right) selected to optimize the number of crimes predicted.

Datos

- Se cuenta con datos de crímenes en Bogot del 16 de abril al 30 de junio de 2017: 16.402 datos.
- Para esta validación se usaron datos de la localidad de Santa Fé, delimitada por latitudes entre $[4.571, 4.629]$ y longitudes > -74.091.

Datos

- Se cuenta con datos de crímenes en Bogot del 16 de abril al 30 de junio de 2017: 16.402 datos.
- Para esta validación se usaron datos de la localidad de Santa Fé, delimitada por latitudes entre $[4.571, 4.629]$ y longitudes > -74.091.

Datos

- 1.676 ($\approx 10\%$) crímenes en la localidad de Santa Fé en el período tratado.

- Se entrenó el modelo de crimen con datos entre 1 y 7 semanas y se validó con las 3 semanas posteriores al entrenamiento, en todos los casos, del 10 al 30 de junio de 2017 (407 crímenes).

Datos

- 1.676 ($\approx 10\%$) crímenes en la localidad de Santa Fé en el período tratado.

- Se entrenó el modelo de crimen con datos entre 1 y 7 semanas y se validó con las 3 semanas posteriores al entrenamiento, en todos los casos, del 10 al 30 de junio de 2017 (407 crímenes).

Validación

- Se divide Bogot ($1.547 km^2$) en 10,946 celdas de $\approx 145 m^2$ cada una; 1.019 celdas en la localidad de Santa F.
- Se entrena el modelo con el correspondiente número de semanas y se predicen los puntos calientes para cada turno de 8 horas, definidos como el 10% de las celdas con mayor probabilidad de crimen.
- Se investigan cuántos crímenes de los datos de validación ocurrieron en los puntos calientes predichos por el modelo.

Validación

- Se divide Bogot ($1.547 km^2$) en 10,946 celdas de $\approx 145 m^2$ cada una; 1.019 celdas en la localidad de Santa F.
- Se entrena el modelo con el correspondiente número de semanas y se predicen los puntos calientes para cada turno de 8 horas, definidos como el 10% de las celdas con mayor probabilidad de crimen.
- Se investigan cuántos crímenes de los datos de validación ocurrieron en los puntos calientes predichos por el modelo.

Validación

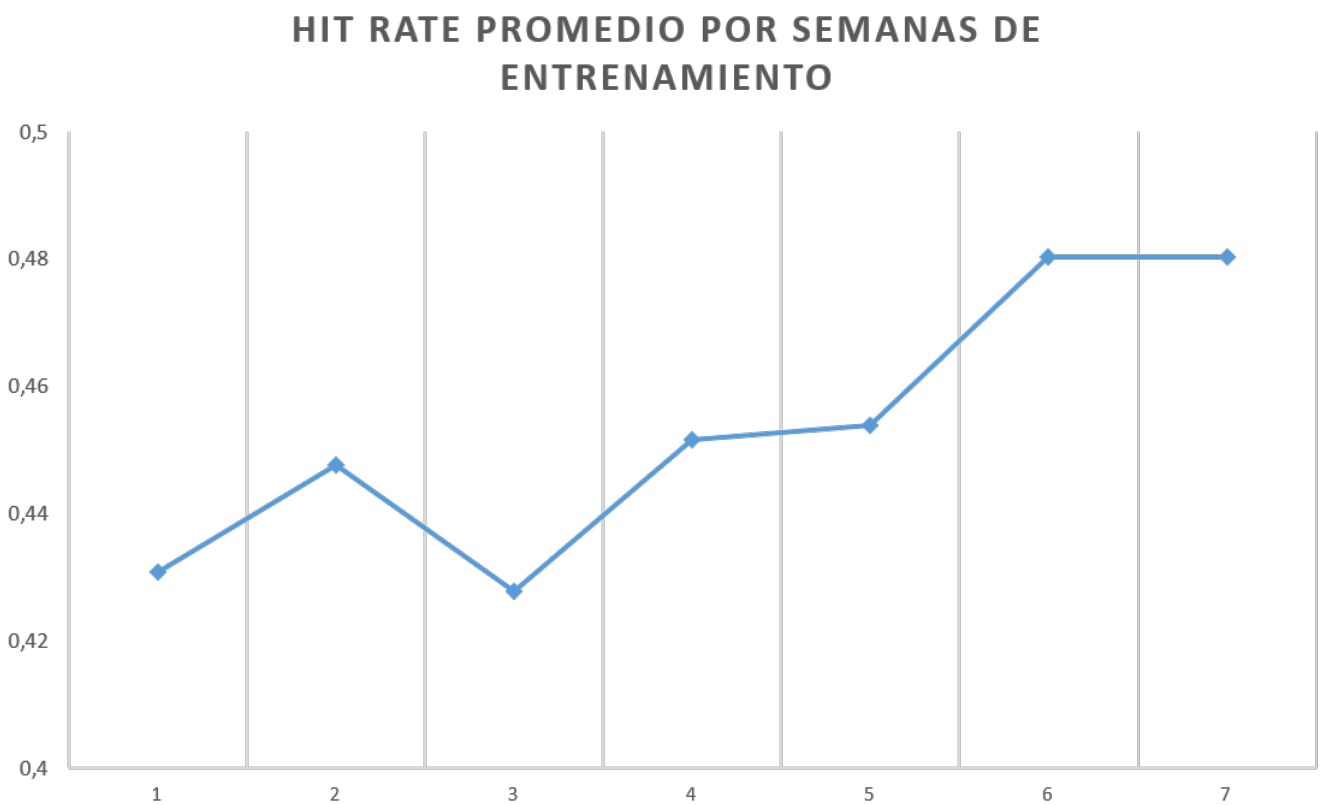
- Se divide Bogot ($1.547 km^2$) en 10,946 celdas de $\approx 145 m^2$ cada una; 1.019 celdas en la localidad de Santa F.
- Se entrena el modelo con el correspondiente número de semanas y se predicen los puntos calientes para cada turno de 8 horas, definidos como el 10% de las celdas con mayor probabilidad de crimen.
- Se investigan cuántos crímenes de los datos de validación ocurrieron en los puntos calientes predichos por el modelo.

Validación: Hit Rate

- La siguiente tabla registra el Hit Rate del modelo al ser entrenado con datos entre 1 y 7 semanas. El Hit Rate se calcul para cada uno de los 21 turnos de 8 horas dentro de cada semana.

Entrenamiento	1	2	3	4	5	6	7
Semana 1	0,49	0,47	0,44	0,44	0,46	0,44	0,44
Semana 2	0,41	0,40	0,40	0,43	0,43	0,46	0,46
Semana 3	0,39	0,48	0,44	0,49	0,48	0,54	0,54
Promedio	0,43	0,45	0,43	0,45	0,45	0,48	0,48

Validación: Hit Rate

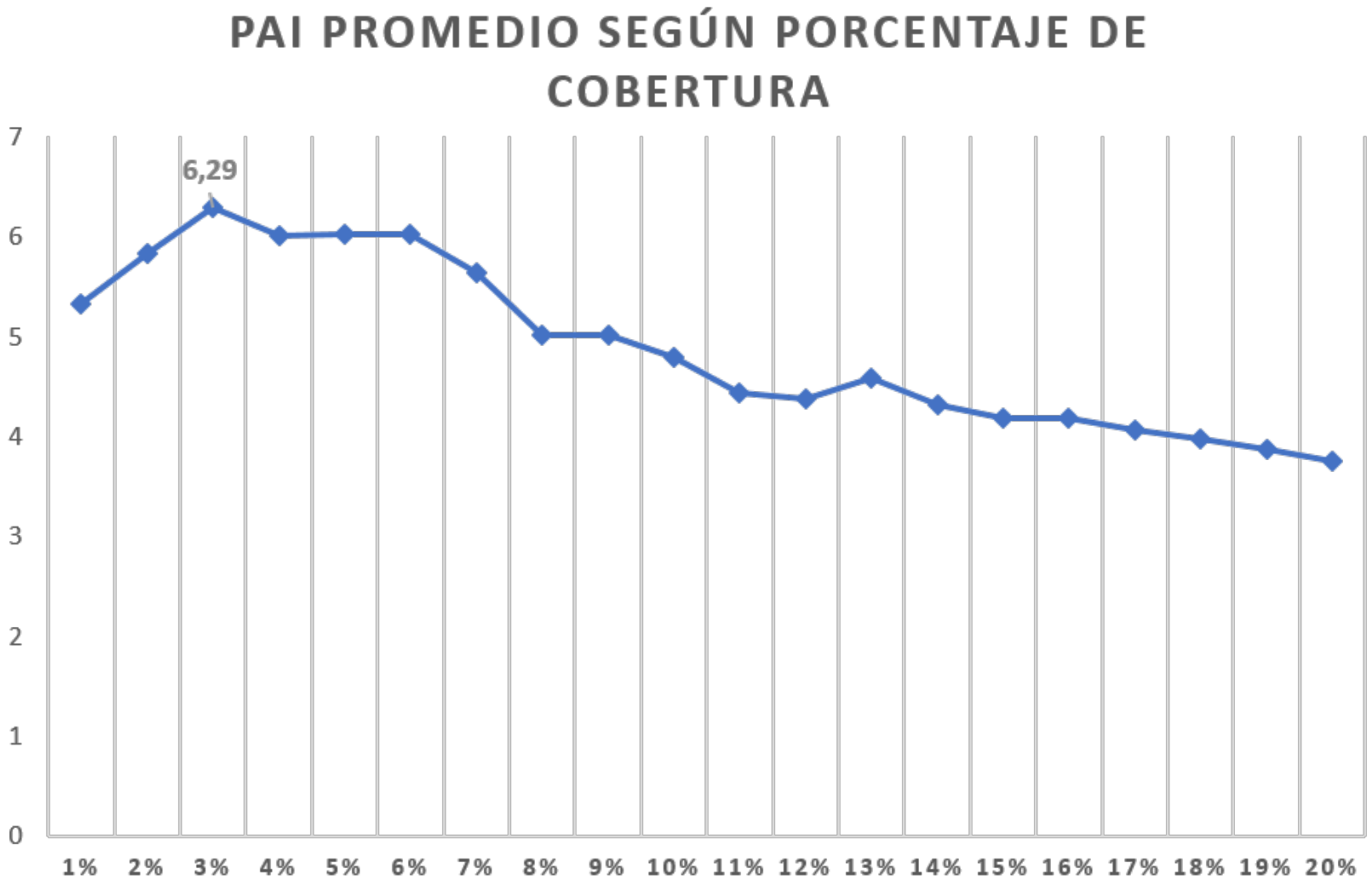


Validación: PAI

Al evaluar la eficiencia del modelo entrenado con 7 semanas, se encontró:

Cobertura	1%	5%	10%	15%	20%
Semana 1	1,53	5,04	4,4	4,21	3,92
Semana 2	4,33	6,14	4,62	3,98	3,64
Semana 3	10,4	6,92	5,43	4,37	3,7
Promedio	5,33	6,02	4,8	4,19	3,75

Validación


PAI PROMEDIO SEGÚN PORCENTAJE DE COBERTURA
6,29
7
6
5
4
3
2
1
0
1%
2%
3%
4%
5%
6%
7%
8%
9%
10%
11%
12%
13%
14%
15%
16%
17%
18%
19%
20%

Validación

- Se puede apreciar que entre mayor sea el número de datos de entrenamiento, el modelo tiene mayor capacidad predictiva.
- Este método de validación se puede utilizar también para calibrar otros parámetros del modelo: tipo de kernel y anchos de banda, pesos otorgados a los datos de entrenamiento, entre otros.

Validación

- Se puede apreciar que entre mayor sea el número de datos de entrenamiento, el modelo tiene mayor capacidad predictiva.
- Este método de validación se puede utilizar también para calibrar otros parámetros del modelo: tipo de kernel y anchos de banda, pesos otorgados a los datos de entrenamiento, entre otros.

Validación

Hit Rate con 7 semanas de entrenamiento y 10% de cobertura de puntos calientes:

Prediccin	bw fijo	bw variable
Semana 1	0,44	0,57
Semana 2	0,46	0,59
Semana 3	0,54	0,62
Promedio	0,48	0,59

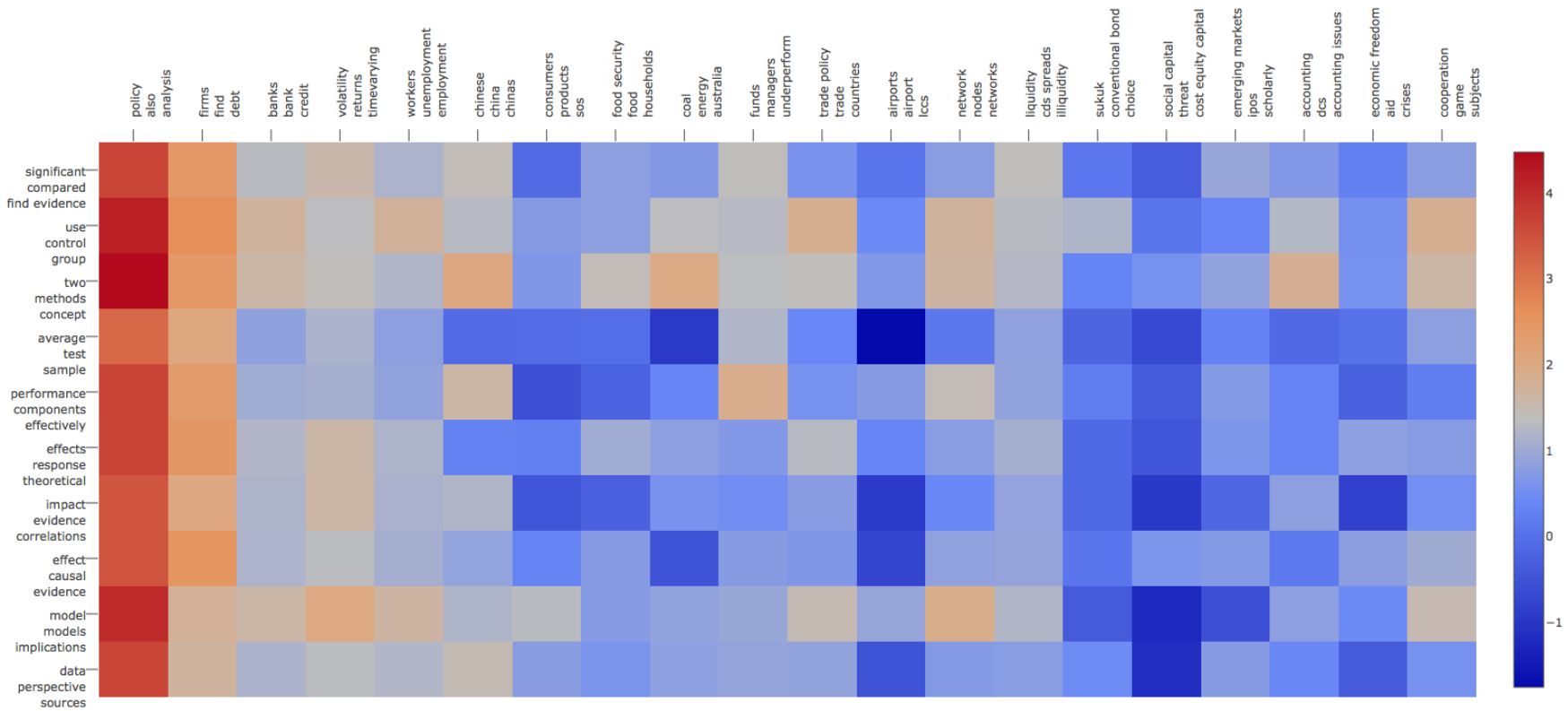
Validación

Hit Rate con 7 semanas de entrenamiento y 10% de cobertura de puntos calientes:

Prediccin	bw fijo	bw variable	KDE
Semana 1	0,44	0,57	0,42
Semana 2	0,46	0,59	0,44
Semana 3	0,54	0,62	0,53
Promedio	0,48	0,59	0,46

Contents

GAP MAPS


policy also analysis
firms find debt
banks bank credit
volatility returns timevarying
workers unemployment employment
chinese china chinas
consumers products sos
food security food households
coal energy australia
funds managers underperform
trade policy trade countries
airports airport lccs
network nodes networks
liquidity cds spreads illiquidity
sukuk conventional bond choice
social capital threat cost equity capital
emerging markets ipos scholarly
accounting dcs accounting issues
economic freedom aid crises
cooperation game subjects
significant compared find evidence
use control group
two methods concept
average test sample
performance components effectively
effects response theoretical
impact evidence correlations
effect causal evidence
model models implications
data perspective sources
4
3
2
1
0
−1

GAP MAPS: Modelo documentos

Definition

Un vocabulario V es una sucesión finita de elementos diferentes $(v_1, \ldots, v_n)$. n es llamado el tamaño del vocabulario V. $v_1, \ldots, v_n$ son llamados los términos del vocabulario V.

Definition

Un documento sobre un vocabulario V es una sucesión $d := (w_1, \ldots, w_n)$ con $w_i \in V$, $\forall i \in \{1, \ldots, n\}$. n es llamado la longitud del documento d.

Definition

Sea V un vocabulario, v_i un término y $d = (w_i, \ldots, w_n)$ un documento sobre este vocabulario.

$$c_i(d) := \sum_{j=1}^{n} \mathbb{I}(v_i = w_j)$$

es la función frecuencia del término v_i

GAP MAPS: Modelo documentos

Definition

El vector de frecuencia de términos de un documento d sobre un vocabulario V es el vector

$$f(d) := (c_1(d), \ldots, c_n(d))$$

Definition

Un Corpus $C = (d_1, \ldots, d_n)$ es una sucesión de documentos sobre un vocabulario V. n es llamado el tamaño del Corpus.

Definition

Dado un Corpus $C = (d_1, \ldots, d_m)$ sobre un vocabulario $V = (v_1, \ldots, v_n)$ se define la matriz de términos y documentos

$$[M_{ij}] = f(d_j)_i, \quad 1 \leq i \leq n, \ 1 \leq j \leq m$$

Esto quiere decir que la entrada en la posición i, j corresponde a la frecuencia con que ocurre el término v_i dentro del documento d_j.

GAP MAPS: Latent Dirichlet Allocation (LDA)

- LDA es un modelo probabilístico generativo de un corpus.
- La idea es que los documentos son representados como mezclas aleatorias sobre tópicos latentes, donde cada tópico es caracterizado como una distribución sobre palabras.

GAP MAPS: Latent Dirichlet Allocation (LDA)

- LDA asume el siguiente proceso generativo para cada documento en un corpus :
  1. Escoja $N \sim \text{Poisson}(\xi)$.
  2. Escoja $\theta \sim \text{Dirichlet}(\alpha)$.
  3. Para cada una de las N palabras w_n:
     - Escoja un tópico $z_n \sim \text{Multinomial}(\theta)$.
     - Escoja una palabra w_n de $p(w_n \mid z_n, \beta)$, una distribución de probabilidad multinomial condicionada en el tópico z_n.

GAP MAPS: Latent Dirichlet Allocation (LDA)

- Las probabilidades de cada palabra estan parametrizadas por una matrix β de dimensiones $k \times V$, donde $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$. Estos son unos valores fijos a ser estimados.

GAP MAPS: Latent Dirichlet Allocation (LDA)

- Dados los parámetros α y β, la función de probabilidad conjunta de una mezcla de tópicos θ, un conjunto de N tópicos , y un conjunto de N palabras está dado por:

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta), \quad (4)$$

donde $p(z_n \mid \theta)$ es θ_i para el único i tal que $z_n^i = 1$.

- Integrando sobre θ y sumando sobre z, obtenemos la distribución marginal de un documento:

$$p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta. \quad (5)$$

GAP MAPS: Latent Dirichlet Allocation (LDA)

- Por último, tomando el producto de las probabilidades marginales de los documentos, obtenemos la probabilidad de un corpus:

$$p(c \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d($$

GAP MAPS: Latent Dirichlet Allocation (LDA)

- Esta distribución es intractable, en general. A pesar de esto diferentes algoritmos de inferencia aproximada pueden ser considerados para LDA, como *Variational Approximation* y *Markov Chain Monte Carlo*.

Automatic summarization

- Cortesía Quantil (proyecto interno).

Contents

Health Records

- Colombia has a mandatory competitive health insurance system.
- Payments are determined using information provided by HMOs.
- Health records: 400 million data items per year, 20 million individuals.
- Potential problems: Risk adjustment, solvency, fraud, data manipulation, misreport, errors, etc.

Health Records

- Colombia has a mandatory competitive health insurance system.
- Payments are determined using information provided by HMOs.
- Health records: 400 million data items per year, 20 million individuals.
- Potential problems: Risk adjustment, solvency, fraud, data manipulation, misreport, errors, etc.

Health Records

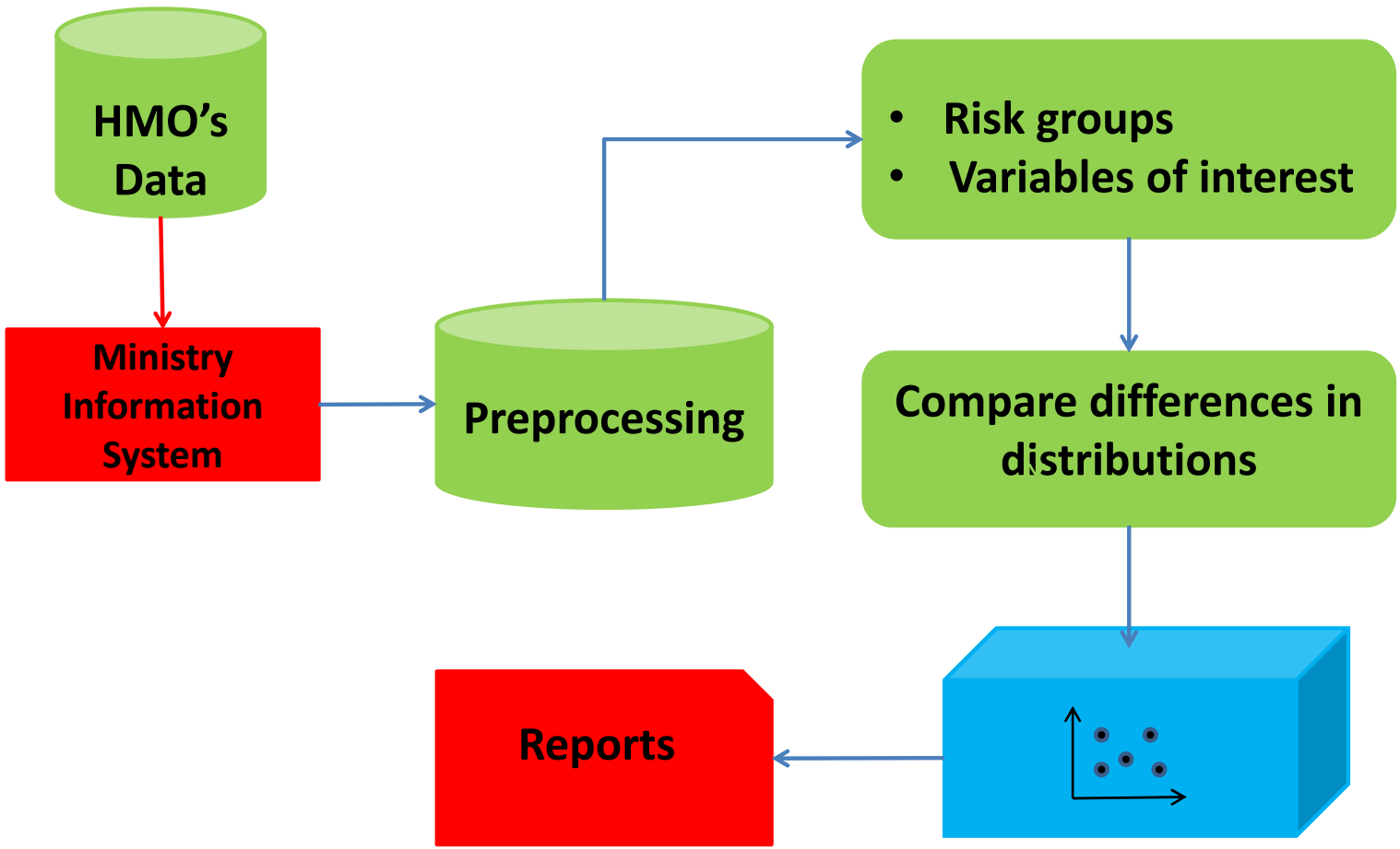
- Colombia has a mandatory competitive health insurance system.
- Payments are determined using information provided by HMOs.
- Health records: 400 million data items per year, 20 million individuals.
- Potential problems: Risk adjustment, solvency, fraud, data manipulation, misreport, errors, etc.

Health Records

- Colombia has a mandatory competitive health insurance system.
- Payments are determined using information provided by HMOs.
- Health records: 400 million data items per year, 20 million individuals.
- Potential problems: Risk adjustment, solvency, fraud, data manipulation, misreport, errors, etc.

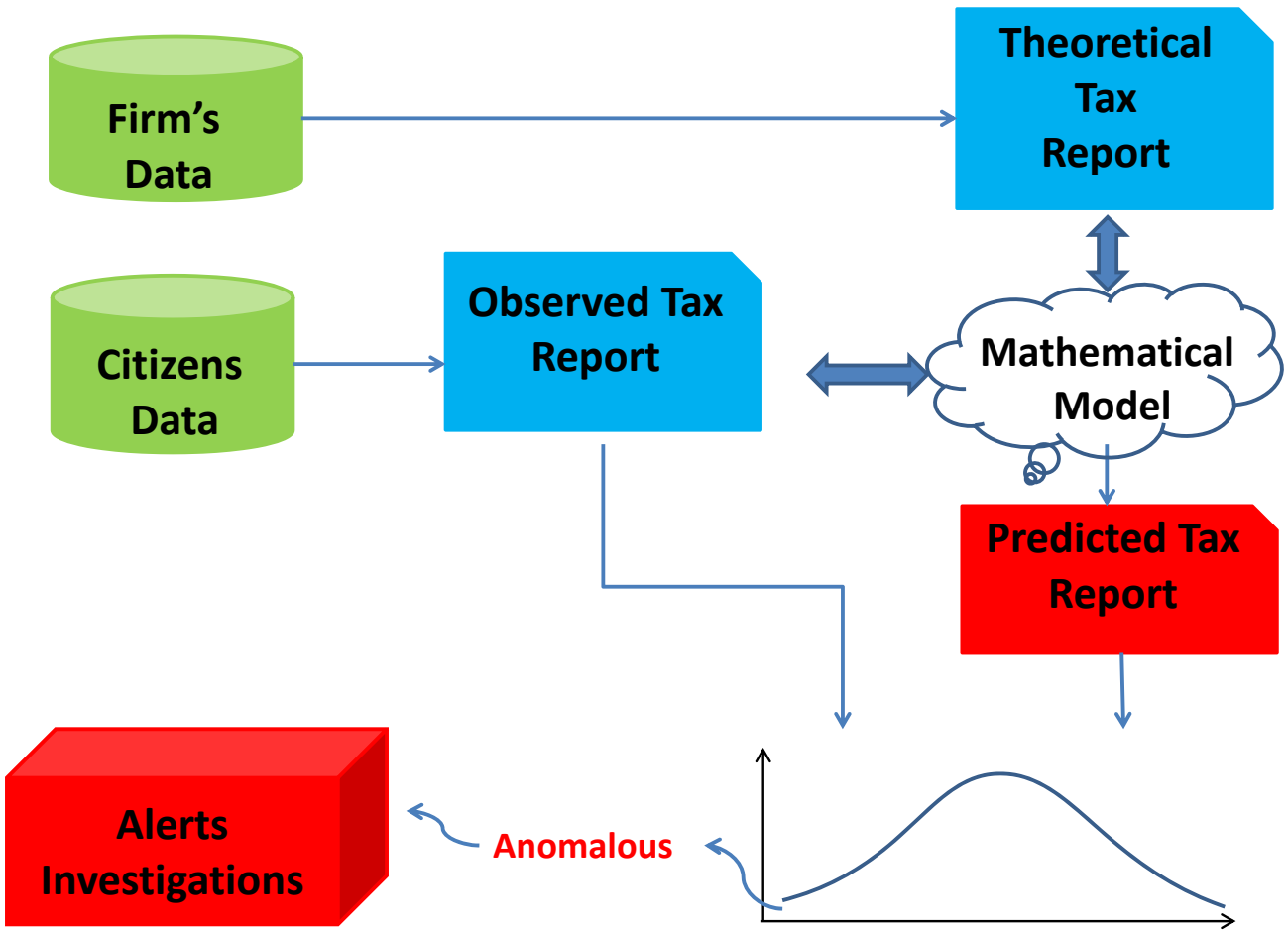
Health Records: Anomaly detection

- Courtesy of Ministry of Health and Social Protection.



Tax Evasion

- Problems: Detect likely cases of tax evasion.



Money Laundry

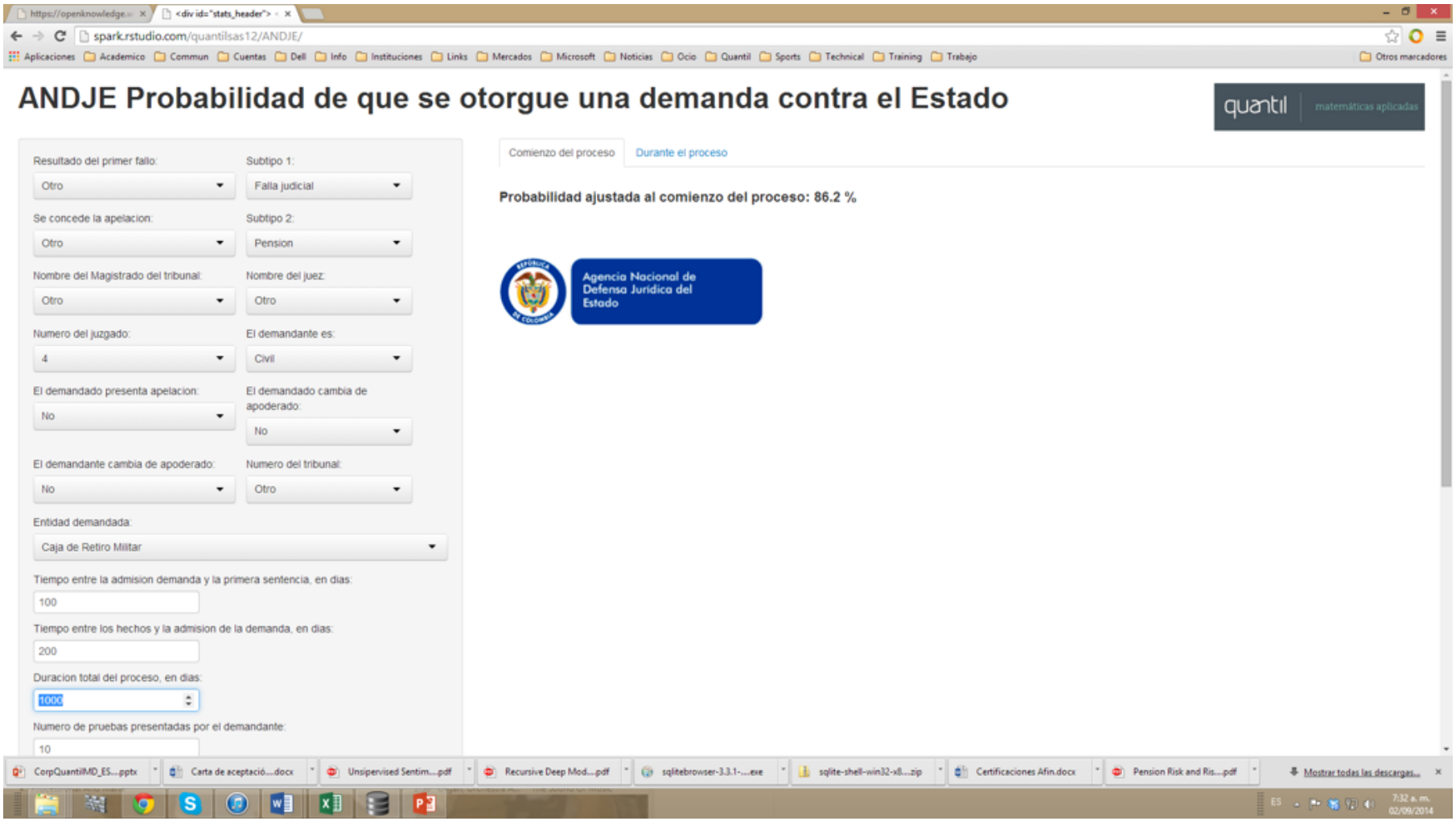
- Problem: Detect anomalous transactions that may be suspicious of money laundry.
- Confidentiality agreements forbids me of explaining further details.

Money Laundry

- Problem: Detect anomalous transactions that may be suspicious of money laundry.
- Confidentiality agreements forbids me of explaining further details.

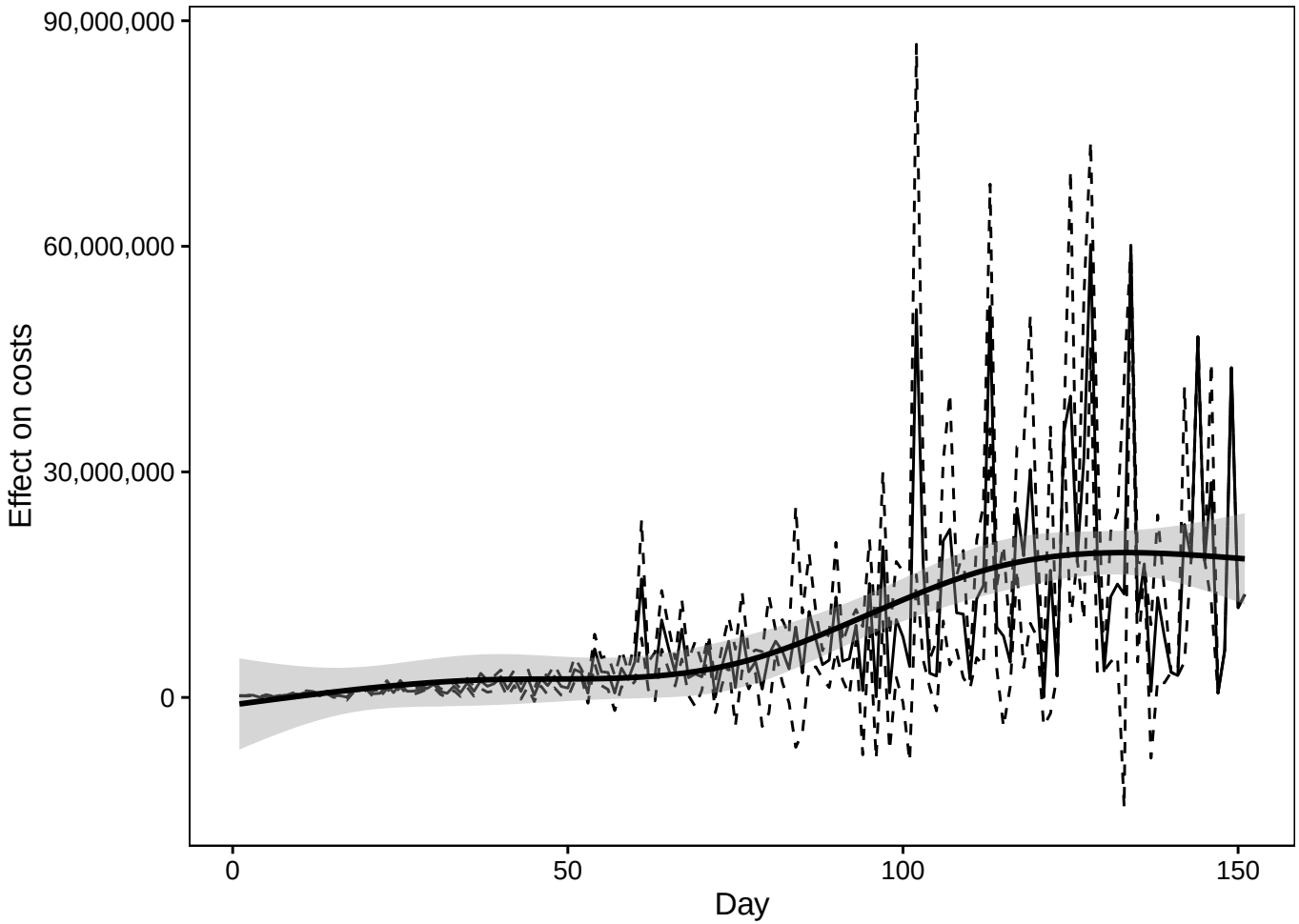
Lawsuits Against the State

- Courtesy of National Agency for Judicial State Defence (ANDJE).



Predicting unnecessary hospitalizations

- Courtesy of Ministry of Health and Social Protection.

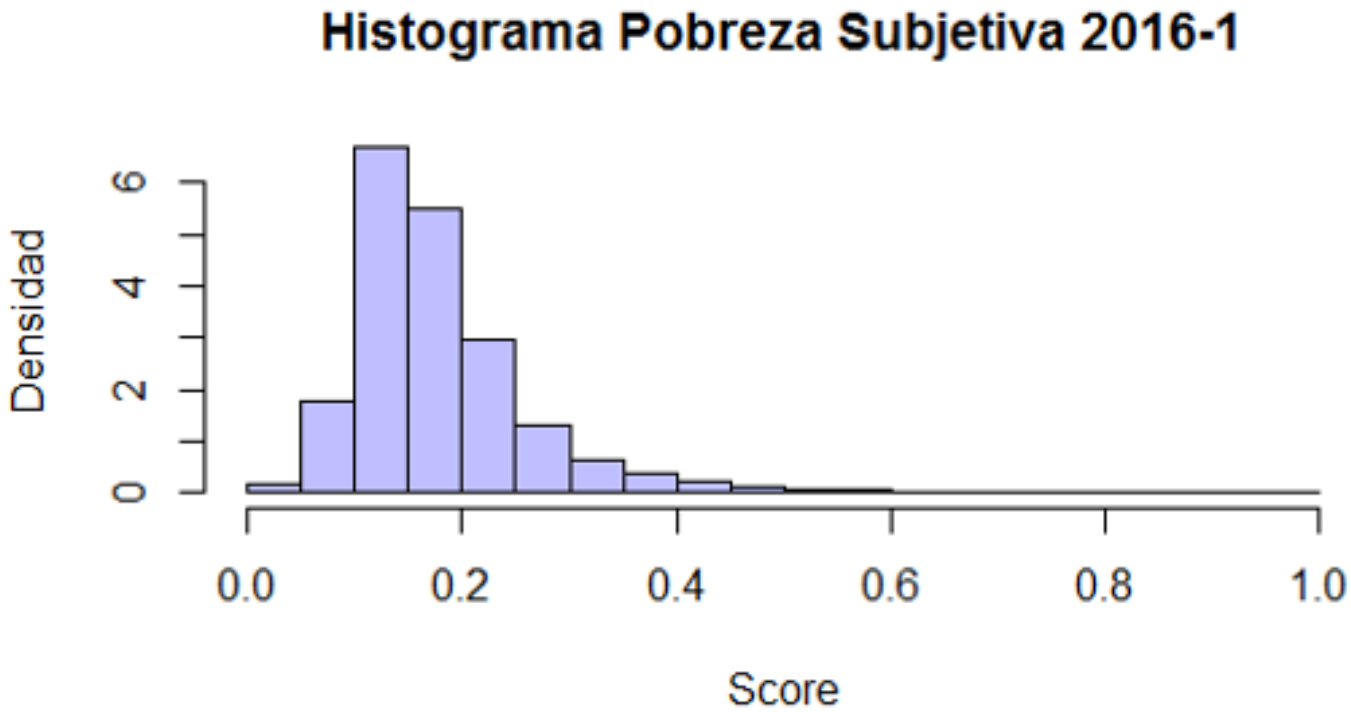


Risk Adjustment

- Courtesy of Ministry of Health and Social Protection.

Subjective Poverty

- Courtesy of National Department of Statistics (DANE).



Gracias!

ariascos@uniandes.edu.co