

# Big Data, Minería y Aprendizaje: Conceptos y Herramientas para Economistas

---

Walter Sosa Escudero

[wsosa@udesa.edu.ar](mailto:wsosa@udesa.edu.ar)

10 de noviembre de 2022

walter sosa escudero

# ¿qué es (y qué no es) la estadística?



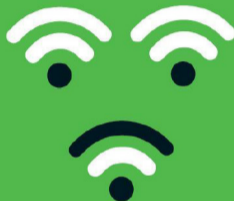
cómo se construyen las predicciones y los datos  
que más influyen en nuestras vidas  
(en medio de la revolución de big data)

 siglo veintiuno  
editores

ciencia que lleva...  
saber mayor

# big data

walter sosa escudero



breve manual para conocer la ciencia de datos  
que ya invadió nuestras vidas

 siglo veintiuno  
editores

ciencia que lleva...  
saber mayor



walter  
sosa  
escudero

# borges, big data y yo

guía nerd  
(y un poco real)  
para perderse  
en el laberinto  
borgesiano

 siglo veintiuno  
editores

ciencia que lleva...  
saber mayor

HILDEGART AHUMADA / M. FLORENCIA GABRIELLI  
MARCOS HERRERA / WALTER SOSA ESCUDERO

# Una nueva econometría

Automatización, big data,  
econometría espacial y estructural

SERIE EXTENSIÓN  
EDUCACIÓN CIENTÍFICA Y TECNOLÓGICA



3. Especificaciones espaciales alternativas	120
3.1. Modelos espaciales de corte transversal	121
3.2. Modelos espaciales en datos de panel	126
3.3. Cuestiones inferenciales y asimóticas	131
4. Interpretación de los modelos con rezago espacial endógeno	135
5. Aplicaciones empíricas	140
5.1. Modelo de crecimiento (Tzur y Koch, 2007)	140
5.2. Consumo de tabaco en datos de panel	145
6. Comentarios finales	148
<b>Capítulo 5. Big data y aprendizaje automático: Ideas y desafíos para economistas</b>	<b>157</b>
Walter Sosa Escudero	
1. Introducción	157
2. Aprendizaje, estadística y econometría	159
2.1. El paradigma inferencial frecuentista en econometría	161
2.2. Aprendizaje automático y construcción de modelos	163
3. Regresión para la predicción	166
3.1. Estimar vs. construir modelos	166
3.2. Complejidad y elección de modelos	168
3.3. Validación cruzada	169
3.4. Regularización: LASSO y ridge	172
4. Clasificación	178
4.1. Riesgo, probabilidades y clasificación	178
4.2. Regresión logística y alternativas	179
4.3. Categorías múltiples	181
4.4. Árboles	182
5. Aprendizaje no supervisado: clusters y dimensionalidad	186
5.1. Clusters	187
5.2. Reducción de dimensionalidad	190

XIII

Advanced Studies in Theoretical and Applied Econometrics 53

Felix Chan  
László Mátyás Editors

# Econometrics with Machine Learning

Springer

Contents	ix
8.2.2. Vectorial Regression	274
8.2.3. CLM	276
8.2.4. Solution Techniques	280
8.3. Geopark Models in the Context of Finance	282
8.3.1. The No-Smart-Sale Environment and Strategy	287
8.3.2. The A-Norm Environment and Strategy	279
8.3.3. Classical Geopark Models for Finance	272
8.3.4. Augmented Geopark Models for Finance Applications	273
8.4. Geopark Models in the Context of Econometrics	278
8.4.1. Formal Contributions	278
8.4.2. Vector Autoregressive Models	280
8.5. Further Integration of Geopark Models with Machine Learning	283
References	285
<b>9. Poverty, Inequality and Development Studies with Machine Learning</b>	<b>290</b>
Walter Sosa Escudero, María Victoria Anselmi and Wendy Blair	
9.1. Introduction	290
9.2. Measurement and Processing	293
9.2.1. Combining Statistics to Improve Data Reliability	294
9.2.2. More Granular Measurements	296
9.2.3. Dimensionality Reduction	296
9.2.4. Data Imputation	298
9.2.5. Methods	307
9.3. Causal Inference	307
9.3.1. Heterogeneous Treatment Effects	307
9.3.2. Optimal Treatment Assignment	312
9.3.3. Handling High Dimensional Data and Debiased ML	313
9.3.4. Machine Learning Counterfactuals	313
9.3.5. New Data Sources for Economics and Treatment	314
9.3.6. Combining Observational and Experimental Data	319
9.4. Computing Power and Tools	326
9.5. Concluding Remarks	327
References	329
<b>10. Machine Learning for Asset Pricing</b>	<b>337</b>
Janja Vukobrat	
10.1. Introduction	337
10.2. How Machine Learning Techniques Can Help Identify Stocking Decision Factors	340
10.3. How Machine Learning Techniques Can Testify to the Asset Pricing Models	345
10.4. How Machine Learning Techniques Can Estimate Linear Factor Models	348
10.4.1. Capital Asset Pricing and Sharpe's (2016) Factor Pricing	348
10.4.2. A New Approach for Estimating Linear Factor Models	349

**WIRED**

GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIGN OPINION

SCIENCE : DISCOVERIES 

## The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson  06.23.08



YouTube AR

Buscar

# 2014 Significance Lecture

## The Big Data Trap

Tim Hardford, Economist, journalist and broadcaster  
Chair: Brian Tarran, Editor, *Significance*

ROYAL  
STATISTICAL  
SOCIETY  
DATA | EVIDENCE | DECISIONS

1:36 / 58:08

RSS 2014 Significance Lecture - The Big Data trap

# Big data: ¿Otra vez arroz?

## DEBATE

2 opiná

142 shares

11

131

3\*

+

**Walter Sosa  
Escudero**  
Profesor  
Asociado, Udesa

Creo conservar, en algún recóndito lugar de mi casa, mi paleta de paddle de cuando en los noventa pensaba que el juego del presente se transformaría en el deporte del futuro. También disfruto de los vinilos de mi adolescencia que escucho casi a diario. Y por alguna razón exótica guardo celosamente una caja de diskettes de mis comienzos con la computación personal, allá en los ochenta.

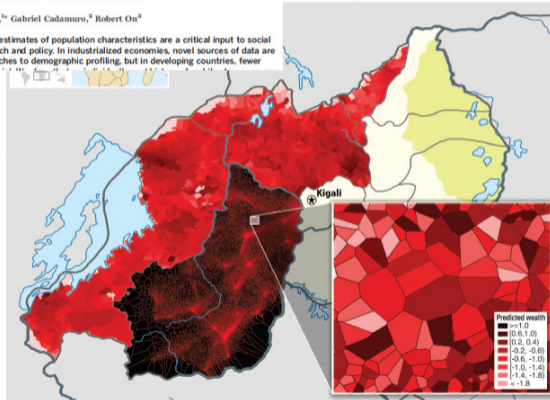
# Pobreza en Rwanda (predecir)

ECONOMICS

## Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,<sup>1\*</sup> Gabriel Cadamuro,<sup>2</sup> Robert On<sup>3</sup>

Accurate and timely estimates of population characteristics are a critical input to social and economic research and policy. In industrialized economies, novel sources of data are enabling new approaches to demographic profiling, but in developing countries, fewer



# Precios en Argentina (medir)



Contents lists available at ScienceDirect

Journal of Monetary Economics

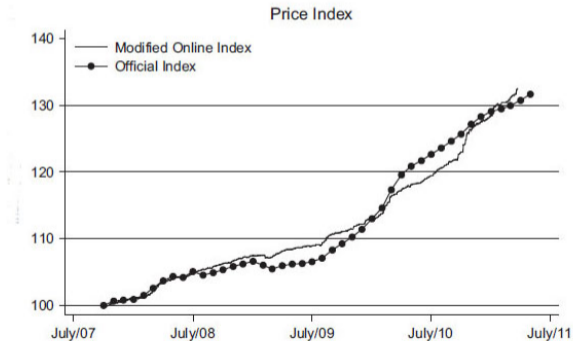
journal homepage: [www.elsevier.com/locate/jme](http://www.elsevier.com/locate/jme)



## Online and official price indexes: Measuring Argentina's inflation

Alberto Cavallo\*

Massachusetts Institute of Technology, Sloan School of Management, 77 Massachusetts Ave 510-512, Cambridge, MA 02139, USA





## **Sales Taxes and Internet Commerce**

Liran Einav

Dan Knoepfle

Jonathan Levin

Neel Sundaresan

AMERICAN ECONOMIC REVIEW  
VOL. 104, NO. 1, JANUARY 2014  
(pp. 1-26)

## Small vs. Big data

### Small data (estadística clásica)

- Extraer lo máximo de **pocos** datos
- Solución: **estructura** (muestreo, modelo)
- Enfoque: muestreo complejo aproxima muestreo al azar (**lento** y **caro**, pero bueno). Teoría, experimentos.

### Big data

- **Muchos** datos (Volumen)
- Muchos datos **no estructurados** (Variedad)
- Muchos datos no estructurados e **inmediatos** (Velocidad)
- 'Condicionablemente baratos'.

$$Y = f(X) + u$$

- Interes en  $f(\cdot)$ . Efecto causal
- Modelo: Teoria, experimento.
- Probabilidades (error estandar, tests)
- Bueno?: insesgado, varianza minima, inferencia valida.

$$Y = f(X) + u$$

- Interes en  $Y$ : predecir, clasificar, medir.
- Modelo: modelo?. Lo **aprendemos**.
- Prediccion puntual (no inferencia).
- Bueno?: Performance predictiva fuera de la muestra.

## Ejemplo: Ridge / LASSO

$$L(\beta) = \sum (y_i - x_i\beta)^2 + \lambda \beta^2$$

- $\lambda = 0$ : MCO.
- $\lambda \neq 0$ : sesgado pero...
- ... **siempre** le puede ganar a MCO en terminos de prediccion.
- Sesgo: pecado mortal.
- **ML**: sesgo puede reducir la varianza dramaticamente.
- Puede lidiar con  $K > n$ .

Esto es ridge. LASSO reemplaza  $\beta^2$  por  $|\beta|$ .

Para  $\lambda \geq 0$  dado, consideremos la siguiente función objetivo (a minimizar):

$$R_l(\beta) = \sum_{i=1}^n (y_i - x_i' \beta)^2 + \lambda \sum_{s=2}^p |\beta_s|$$

(el primer coeficiente corresponde al intercepto).

- ¿Si  $\lambda = 0$ ?
- ¿Si  $\lambda = \infty$ ?
- $\sum_{i=1}^n (y_i - x_i' \beta)^2$  penaliza falta de ajuste.
- ¿  $\sum_{s=2}^p |\beta_s|$  ?

$$R_I(\beta) = \sum_{i=1}^n (y_i - x_i' \beta)^2 + \lambda \sum_{s=2}^p |\beta_s|$$

- LASSO magic: automáticamente elige que variables entran ( $\beta_s \neq 0$ ) y cuales no ( $\beta_s = 0$ )
- Por que? Coeficientes anulados como 'soluciones de esquina'
- $R_I(\beta)$  es una función no diferenciable.

Caso super simple

$$R_I(b) = \sum_{i=1}^n (y_i - x_i b)^2 + \lambda |b|$$

$$R_I(b) = SRC(b) + \lambda |b|$$

Supongamos  $\hat{b} > 0$ . Empecemos en  $b = 0$

- Ingreso Mg:  $-\frac{d SRC(b)}{db} \Big|_{b=0}$  (positivo y decreciente)
- Costo Mg:  $\lambda$  (positivo y constante)



$$R_l(b) = SRC(b) + \lambda|b|$$

- **LASSO**: poner variables solo si son suficientemente relevantes.
- En general,  $\hat{b}_l = 0$  para variables irrelevantes, y  $\hat{b}_l$  esta 'corrido hacia cero' para las relevantes.
- Shrinkage: estimacion *sesgada* por la regularizacion.
- Regularizar: 'disciplinar' el modelo hacia la hipotesis nula de no significatividad.
- Por que? Eliminar variables induce sesgo pero puede bajar dramaticamente la varianza, mejora ECM.

# Ejemplo: Hitters

Hitters

## Format

A data frame with 322 observations of major league players on the following 20 variables.

AtBat Number of times at bat in 1986

Hits Number of hits in 1986

HmRun Number of home runs in 1986

Runs Number of runs in 1986

RBI Number of runs batted in in 1986

Walks Number of walks in 1986

Years Number of years in the major leagues

CAtBat Number of times at bat during his career

CHits Number of hits during his career

CHmRun Number of home runs during his career

CRuns Number of runs during his career

CRBI Number of runs batted in during his career

CWalks Number of walks during his career

League A factor with levels A and N indicating player's league at the end of 1986

Division A factor with levels E and W indicating player's division at the end of 1986

PutOuts Number of put outs in 1986

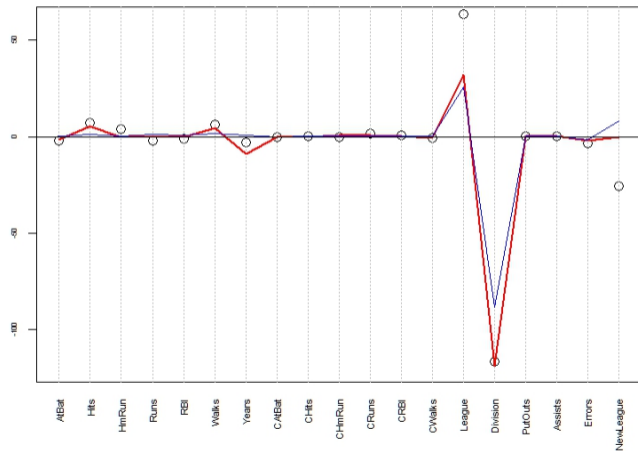
Assists Number of assists in 1986

Errors Number of errors in 1986

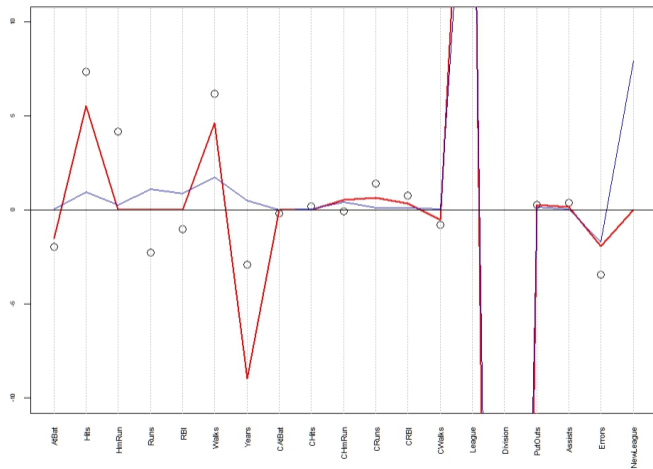
Salary 1987 annual salary on opening day in thousands of dollars

NewLeague A factor with levels A and N indicating player's league at the beginning of 1987

# Ejemplo: Hitters



# Ejemplo: Hitters



- Etiqueta estadística: ex-ante. Teoría, identificación 'limpia' (consistencia). Inferencia robusta.
- Machine learning: ex-post, iterativa. Cross validation.
- Machine learning **construye** el modelo más que lo estima, en base a la performance predictiva **fuera de la muestra**. Adios al  $R^2$  (y a MCO? Mmm...).

- Dependencias (realmente tenemos big data?. Trump effect)
- Choice based sampling.
- Contracticos (podemos tener *todos* los datos?).
- Falacia de la correlacion.
- Transparencia / privacidad.
- Comunicabilidad. Caja negra (deep learning, forests, etc.)
- Consenso social/politico.

- More mas que big.
- Complejidad, heterogeneidad. No linealidades. Maldicion de la dimensionalidad.
- Rapido (crucial para la politica). Google Flu Trends. Price scrapping.
- Oportunidad para diseño de experimentos.
- Complementa a las estadísticas oficiales (no reemplaza).
- Cobertura. Rural vs. urbano, etc..