Lecture 1: Introducción

Big Data and Machine Learning en el Mercado Inmobiliario
Educación Continua

Ignacio Sarmiento-Barbieri

Universidad de los Andes

August 17, 2021

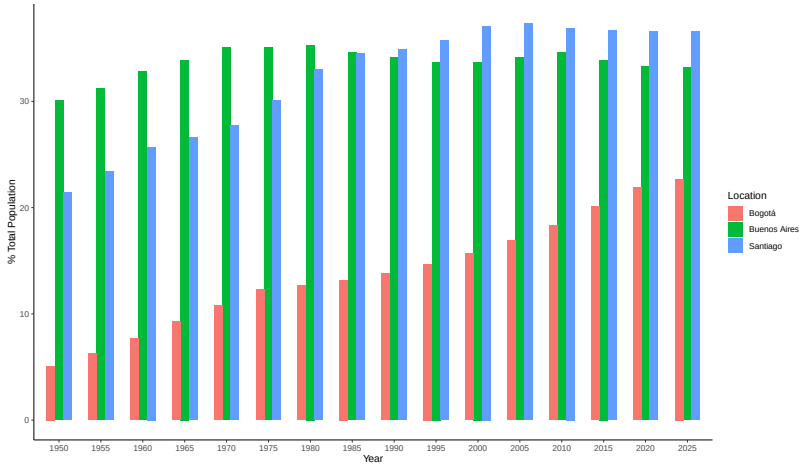
Agenda

Motivación

- Algunas Cifras
  - El 33% de la población de Argentina vive en Bs. As. ($< 1\%$ del territorio)
  - El 36% de la población de Chile vive en Santiago ($\sim 2\%$ del territorio)
  - El 20% de la población de Colombia vive en Bogotá ($\sim 7\%$ del territorio)
  - El 70% de la población de EEUU viven en Ciudades ($\sim 4\%$ del territorio)

Motivación

Figure 1: Porcentaje de población



Fuente: https://population.un.org/

Motivación

Table 1

City	Price to Income Ratio	Mortgage as % of Income	Price to Rent Ratio City Centre	Price to Rent Ratio Outside of Centre
Buenos-Aires	32.46	1,466.89%	37.40	43.65
Santiago	19.80	144.36%	27.54	24.40
Bogota	22.84	292.48%	19.65	21.46
New-York	9.27	64.65%	22.16	18.84

Fuente: https://www.numbeo.com/property-investment

Big Data and Machine Learning

- ¿Qué es Big Data (las 3 V's) ?
  - **V**olumen (n y k)
  - **V**ariedad
  - **V**elocidad
- ¿Machine Learning?
  - Predicción Robusta fuera de muestra
  - $\neq$ Estadística Clásica (Small Data?, Inferencia)

Ejemplo Abstracto

$$y_i \approx \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} \tag{1}$$

- Estadística/Econometría $E(\hat{\beta}_j) = \beta_j$
- Machine learning:

$$y_{n+1} \approx \beta_1 x_{1n+1} + \beta_2 x_{2n+1} + ... + \beta_k x_{kn+1} \tag{2}$$

- hacer que $\hat{y}_{n+1}$ sea lo mas cercano posible a y_{n+1}

Ejemplo Concreto

La primera victoria y derrota del Big Data y Machine Learning

- Contexto ¿similar? al de hoy: Epidemia de la gripe A en 2009
- En EEUU la forma de monitorear es a través de reportes de la CDC
- La CDC agrega a nivel de ciudad, condado, estado, región y a nivel nacional
- Todo esto llevaba aproximadamente 10 días $\rightarrow$ demasiado tiempo para una epidemia

Ejemplo Concreto

Google se ha unido a la conversación

- Google propuso un mecanismo ingenioso: **Google Flu Trends**
- Punto de partida:
  - Proporcion de visitas semanales por Gripe A en hospitales
  - 9 regiones $\times$ 5 años (2003-2007) = $2,340$ datos
  - Estos son los datos que tomaban 10 dias en elaborarse (comparemos con la Colombia de 2009)
- Google cruzó estos datos con las busquedas sobre la gripe A
- Con estos datos, construyeron un modelo para predecir intensidad de gripe A

Ejemplo Concreto

Google se ha unido a la conversación

- Un solo modelo?
- Los investigadores de Google estimaron **450 millones** de models
- Eligieron el que mejor predice sobe la intensidad de busqueda
- Les permite tener informacion diaria, semanal o mensual para cualquier punto de EEUU y el mundo
- A Google le toma 1 día lo que a la CDC 10!

Ejemplo Concreto

Google se ha unido a la conversación

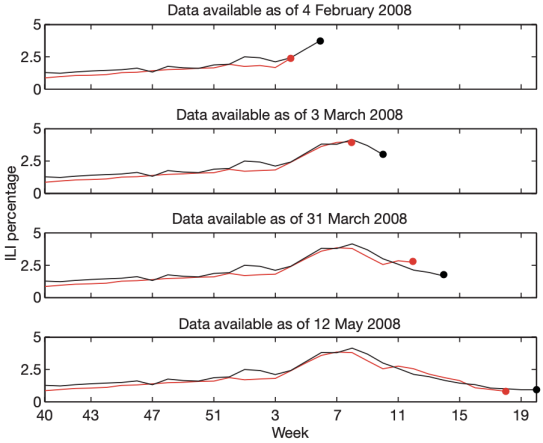


Figure 3 | ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5 we detected a sharply increasing ILI percentage in the mid-Atlantic region; similarly, on 3

Ejemplo Concreto

El rey ha muerto, larga vida al rey

- Qué tienen en común Google Flu y Elvis?
  - Abanderados de la revolución
  - Definió y redefinió las reglas sistemáticas para hallar la solución a un problema
  - Éxito rotundo $\rightarrow$ Publicacion en Nature!
    https://www.nature.com/articles/nature07634
  - Pero como a Elvis el éxito fue efímero
  - La predicciones comenzaron a sobre-estimar
  - Google Flu esta ahora archivado (disponible al publico)
  - Continúa recolectando datos pero solo algunas instituciones científicas tienen accesso

La crítica de Lucas

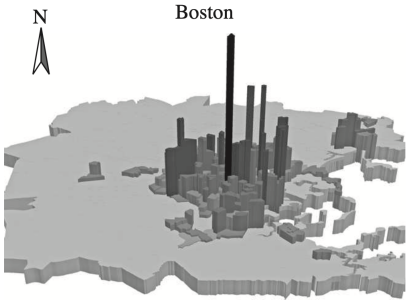
"Given that the structure of an econometric model consist of optimal decision rules of economic agents, and that optimal decision rule vary systematically with change in the structure of series relevant of the decision maker, it follows that any change in policy will systematically alter the structure of econometric models"

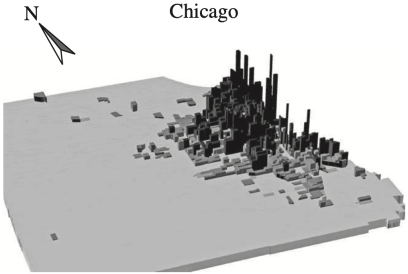
Lucas, 1976

Economía Urbana y este curso

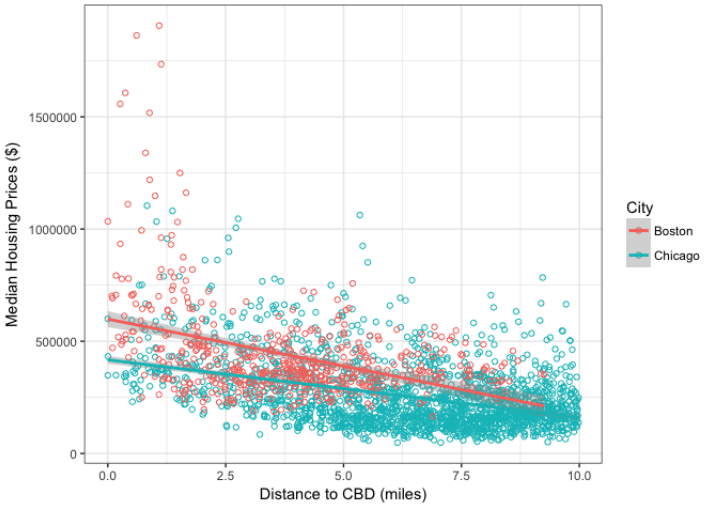
- El economista urbano busca formular de una explicación económica rigurosa para entender regularidades observadas en las estructuras espaciales de las ciudades del mundo real.
- El más obvio entre ellos es la dramática variación espacial en la intensidad del uso del suelo urbano.
- Nuestro 'swiss army knife' es el concepto de equilibrio espacial
- En este curso la teoría nos va a guíar y usaremos herramientas Big Data y Machine Learning para entender patrones de las ciudades y el precio de las propiedades.

Economía Urbana y este curso





Economía Urbana y este curso



Organización del curso

- Clases: teoría + Práctica en `R` y/o `Python`
- Primera mitad del curso (5 clases) enfocado en `R`
  - Datos Espaciales en `R`
  - APIs
- Segunda mitad del curso (5 clases) enfocado en `Python`
  - Webscraping
  - Machine Learning models
- Certificado de participación a los estudiantes que cursen como mínimo el 85% de las sesiones (9/10)

Para seguir leyendo

- Glaeser, E. L. (2008). Cities, agglomeration, and spatial equilibrium. Oxford University Press.
- Lucas, Robert Jr, (1976). Econometric policy evaluation: A critique. Carnegie-Rochester Conference Series on Public Policy, Elsevier, vol. 1(1), pages 19-46, January.
- O'Sullivan, A. Urban Economics. 8va Edición.
- Taddy, M. (2019). Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions. McGraw Hill Professional
- Tom Shaffer The 42 V's of Big Data and Data Science. https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html

Volvemos en 5 min con `R` y `Python`

Details

- Price to Income Ratio is the basic measure for apartment purchase affordability (lower is better). It is generally calculated as the ratio of median apartment prices to median familial disposable income, expressed as years of income (although variations are used also elsewhere). Our formula assumes and uses:
  - net disposable family income, as defined as 1.5 * the average net salary (50% is assumed percentage of women in the workforce) median apartment size is 90 square meters
  - price per square meter (the formula uses) is the average price of square meter in the city center and outside of the city center
- Mortgage as Percentage of Income is a ratio of the actual monthly cost of the mortgage to take-home family income (lower is better). Average monthly salary is used to estimate family income. It assumes 100% mortgage is taken on 20 years for the house(or apt) of 90 square meters which price per square meter is the average of price in the city center and outside of city center.
- Loan Affordability Index is an inverse of mortgage as percentage of income. Used formula is : (100 / mortgage as percentage of income) (higher is better).
- Price to Rent Ratio is the average cost of ownership divided by the received rent income (if buying to let) or the estimated rent that would be paid if renting (if buying to reside). Lower values suggest that it is better to buy rather than rent, and higher values suggest that it is better to rent rather than buy. Our formula to estimate rent per square meter assumes 1 bedroom apt has 50 square meters and 3 bedroom apartment has 110 square meters. It doesn't take into account taxes or maintenance fees.