Mining Administrative Databases: An application to Developing Countries

Alvaro J. Riascos
University of los Andes and Quantil

HACKER DOJO

September 19, 2014

Contents

Introduction

- Traditional databases (transactional, communications, images, financial, auctions, health records, surveys, etc.) contain core information.
- Usually it is costly to obtain this information or there are real consequences at stake.
- Their study is sometimes theory based (auctions, health records, surveys, etc.)
- Machine learning and data mining techniques provide new tools for exploring traditional databases complemented or not with sound theories.

Introduction

- Traditional databases (transactional, communications, images, financial, auctions, health records, surveys, etc.) contain core information.
- Usually it is costly to obtain this information or there are real consequences at stake.
- Their study is sometimes theory based (auctions, health records, surveys, etc.)
- Machine learning and data mining techniques provide new tools for exploring traditional databases complemented or not with sound theories.

Introduction

- Traditional databases (transactional, communications, images, financial, auctions, health records, surveys, etc.) contain core information.
- Usually it is costly to obtain this information or there are real consequences at stake.
- Their study is sometimes theory based (auctions, health records, surveys, etc.)
- Machine learning and data mining techniques provide new tools for exploring traditional databases complemented or not with sound theories.

Introduction

- Traditional databases (transactional, communications, images, financial, auctions, health records, surveys, etc.) contain core information.
- Usually it is costly to obtain this information or there are real consequences at stake.
- Their study is sometimes theory based (auctions, health records, surveys, etc.)
- Machine learning and data mining techniques provide new tools for exploring traditional databases complemented or not with sound theories.

Introduction

- This talk provides a few examples of successful mining of these traditional databases particularly relevant to developing countries.
- All examples are taken from Colombian databases.
- Confidentiality agreements will pose some limits to the discussion.

Introduction

- This talk provides a few examples of successful mining of these traditional databases particularly relevant to developing countries.
- All examples are taken from Colombian databases.
- Confidentiality agreements will pose some limits to the discussion.

Introduction

- This talk provides a few examples of successful mining of these traditional databases particularly relevant to developing countries.
- All examples are taken from Colombian databases.
- Confidentiality agreements will pose some limits to the discussion.

Contents

Health records

- Colombia has a mandatory competitive health insurance system.
- Every person is entitled to a (broad) health insurance package.
- Specialized private institutions (HMOs) compete for affiliates and act as health insurance. They pay health providers (hospitals) for health services.
- In turn, at the beginning of every year the government pays insurance providers a risk adjusted payment per affiliate (a capitation payment).

Health records

- Colombia has a mandatory competitive health insurance system.
- Every person is entitled to a (broad) health insurance package.
- Specialized private institutions (HMOs) compete for affiliates and act as health insurance. They pay health providers (hospitals) for health services.
- In turn, at the beginning of every year the government pays insurance providers a risk adjusted payment per affiliate (a capitation payment).

Health records

- Colombia has a mandatory competitive health insurance system.
- Every person is entitled to a (broad) health insurance package.
- Specialized private institutions (HMOs) compete for affiliates and act as health insurance. They pay health providers (hospitals) for health services.
- In turn, at the beginning of every year the government pays insurance providers a risk adjusted payment per affiliate (a capitation payment).

Health records

- Colombia has a mandatory competitive health insurance system.
- Every person is entitled to a (broad) health insurance package.
- Specialized private institutions (HMOs) compete for affiliates and act as health insurance. They pay health providers (hospitals) for health services.
- In turn, at the beginning of every year the government pays insurance providers a risk adjusted payment per affiliate (a capitation payment).

Health records

- Capitation payments are determined using information provided by HMOs. Risk of manipulation, misreport, errors, etc.
- The problem to be solved is to design a system that helps the Ministry to detect all sorts of anomalous behaviour.
- Key elements:
  1. Diagnostic related groups: Long Duration Medical Conditions Groups - LDG (this is where theory is critical).
  2. Unsupervised learning techniques (we use the concept of relative entropy).

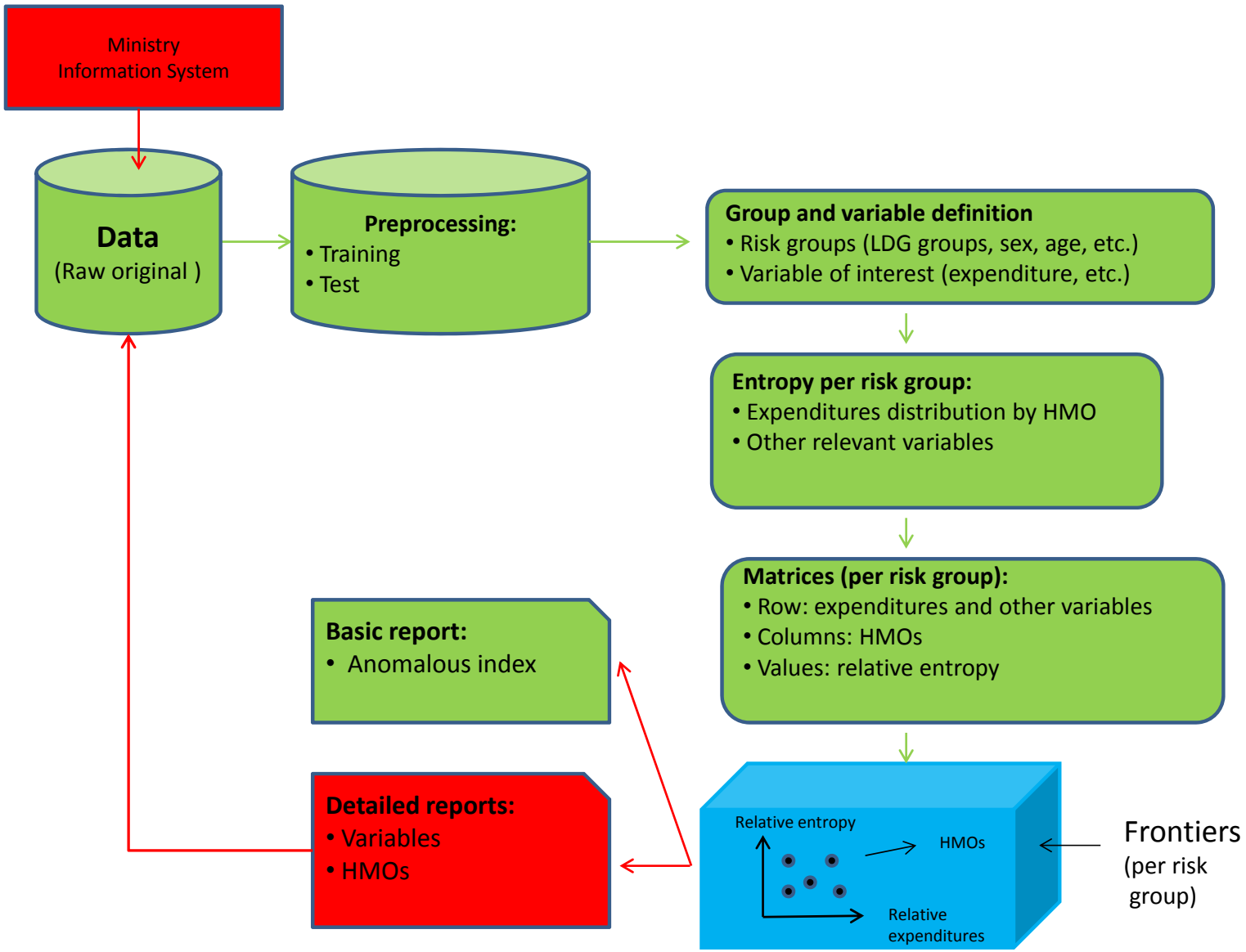
Health records

- Capitation payments are determined using information provided by HMOs. Risk of manipulation, misreport, errors, etc.
- The problem to be solved is to design a system that helps the Ministry to detect all sorts of anomalous behaviour.
- Key elements:
  1. Diagnostic related groups: Long Duration Medical Conditions Groups - LDG (this is where theory is critical).
  2. Unsupervised learning techniques (we use the concept of relative entropy).

Health records

- Capitation payments are determined using information provided by HMOs. Risk of manipulation, misreport, errors, etc.
- The problem to be solved is to design a system that helps the Ministry to detect all sorts of anomalous behaviour.
- Key elements:
  1. Diagnostic related groups: Long Duration Medical Conditions Groups - LDG (this is where theory is critical).
  2. Unsupervised learning techniques (we use the concept of relative entropy).

Health records



Health records

- Any data visualization technique, correlation analysis, etc. on this data set is extremely valuable for policy makers and health scientists.
- We've explored:
  1. Risk adjustment (exante and expost).
  2. Evaluated policy interventions (e.g. the expected cost of structural reforms such as introducing an almost unlimited health benefit package).

Health records

- Any data visualization technique, correlation analysis, etc. on this data set is extremely valuable for policy makers and health scientists.
- We've explored:
  1. Risk adjustment (exante and expost).
  2. Evaluated policy interventions (e.g. the expected cost of structural reforms such as introducing an almost unlimited health benefit package).

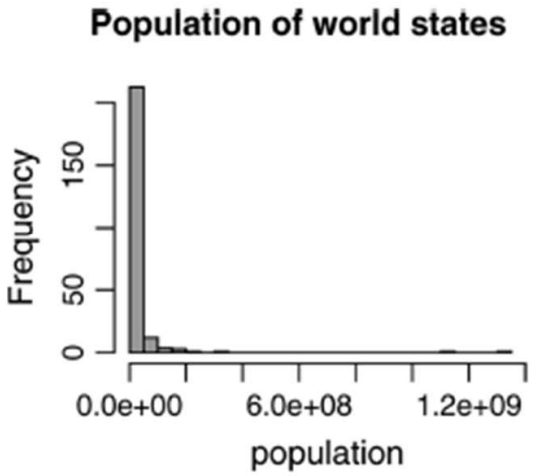
Contents

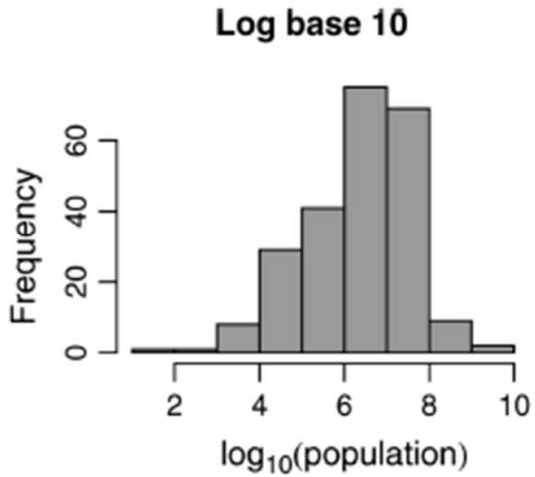
Nationwide Elections

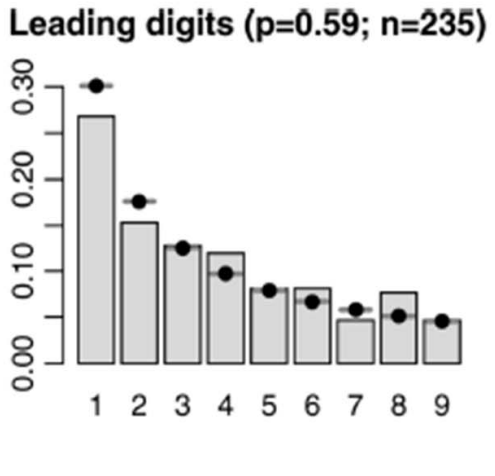
- Problem: detect anomalous behaviour in voting.
- Three main techniques:
  1. Benford Law (little down to earth theory).
  2. Beber and Sccaco tests (some theory).
  3. Supervised learning (a theory of fraud).

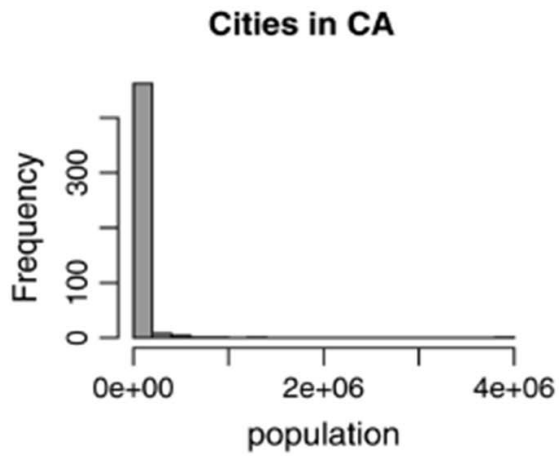
Nationwide Elections

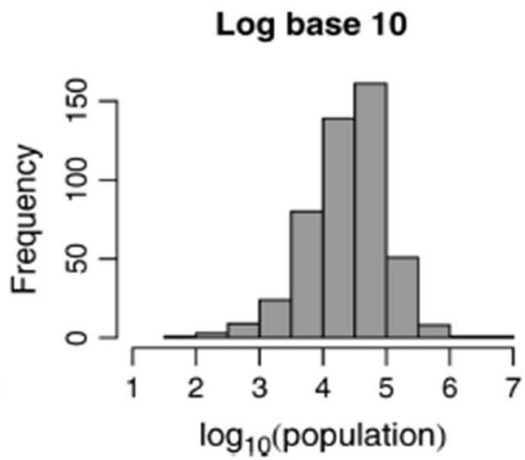
- Benford law

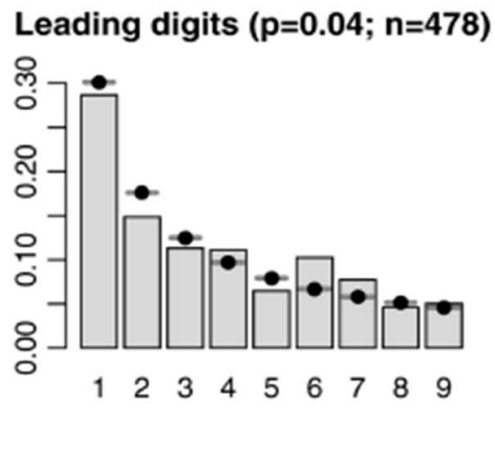












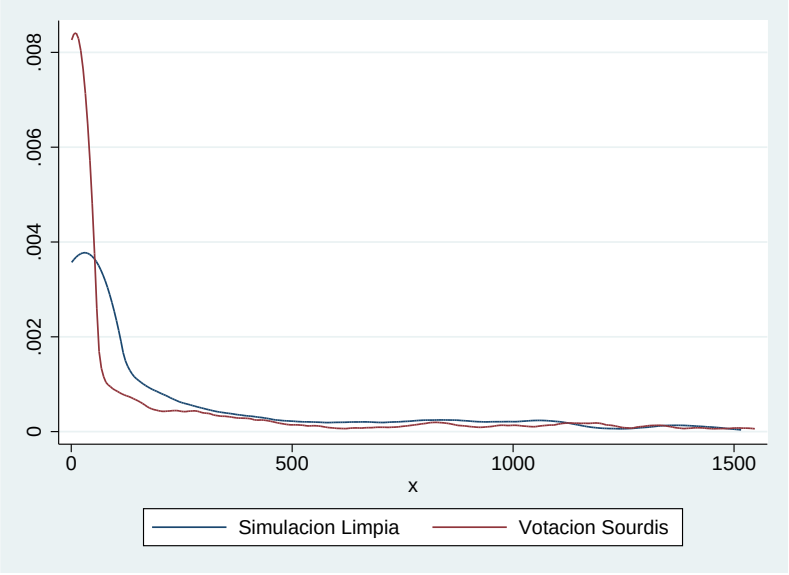
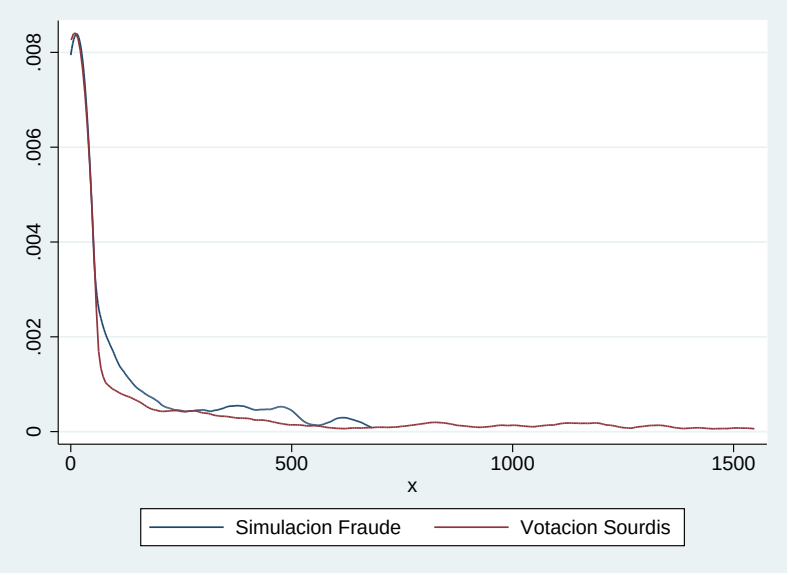
Nationwide Elections

- Beber and Sccaco tests are based on the following facts:
  1. Theory: The last digit, consecutive digits and repeated digits of randomly generated numbers is distributed uniformly.
  2. Evidence: People rarely produce this pattern when asked for random digits.

Nationwide Elections

- Based on Cantu and Saiegh (2010): A Supervised Machine Learning Procedure to Detect Electoral Fraud Using Digital Analysis.
  1. Generate random numbers as if they had been those of a clean election.
  2. Manipulate the random generator using some hypothesis on the way fraud was implemented (ballot stuffing). Calibrate this model to observed distributions.
  3. Train a classification model using only statistics based on digits (last digit, consecutive digits, etc.)
  4. Classify your election.

Nationwide Elections

Simulacion Limpia (Sourdis)	Simulacion Fraude (Sourdis)
	

Contents

Copy Detection in Nationwide Multiple Choice Tests

- In Colombia all high school students take mandatory multiple choice questions at different stages.
- There are all types of incentives for manipulations at student, school or teachers level.
- Coincidence on wrong answers maybe due to chance but only to some extent.
- Performance depends on non observables (ability, teacher quality) and others in principle observable, but in practice not observable (parents education, etc.)
- We estimate an polytomous item response model and do hypothesis testing.

Copy Detection in Nationwide Multiple Choice Tests

- In Colombia all high school students take mandatory multiple choice questions at different stages.
- There are all types of incentives for manipulations at student, school or teachers level.
- Coincidence on wrong answers maybe due to chance but only to some extent.
- Performance depends on non observables (ability, teacher quality) and others in principle observable, but in practice not observable (parents education, etc.)
- We estimate an polytomous item response model and do hypothesis testing.

Copy Detection in Nationwide Multiple Choice Tests

- In Colombia all high school students take mandatory multiple choice questions at different stages.
- There are all types of incentives for manipulations at student, school or teachers level.
- Coincidence on wrong answers maybe due to chance but only to some extent.
- Performance depends on non observables (ability, teacher quality) and others in principle observable, but in practice not observable (parents education, etc.)
- We estimate an polytomous item response model and do hypothesis testing.

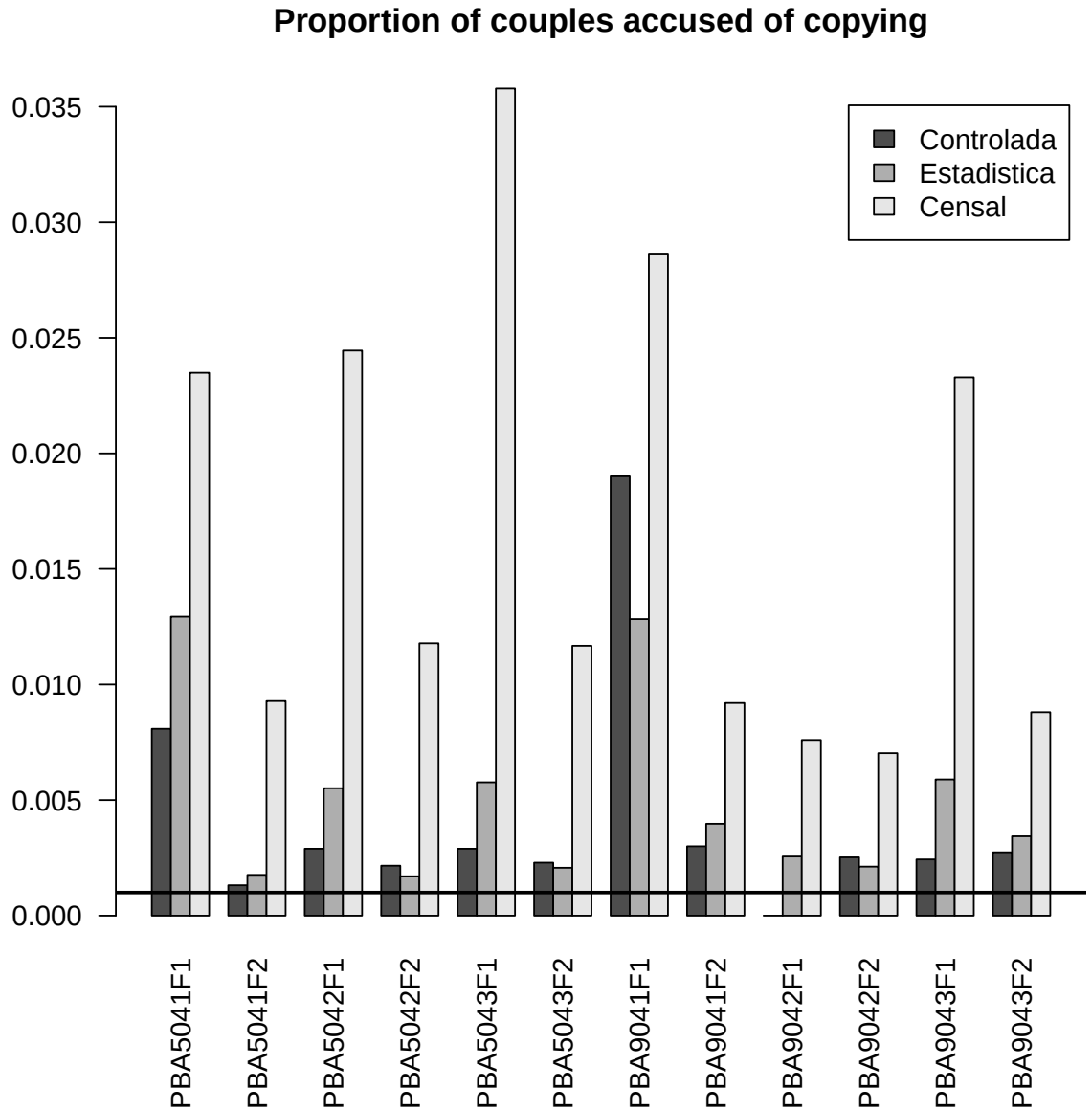
Copy Detection in Nationwide Multiple Choice Tests

- In Colombia all high school students take mandatory multiple choice questions at different stages.
- There are all types of incentives for manipulations at student, school or teachers level.
- Coincidence on wrong answers maybe due to chance but only to some extent.
- Performance depends on non observables (ability, teacher quality) and others in principle observable, but in practice not observable (parents education, etc.)
- We estimate an polytomous item response model and do hypothesis testing.

Copy Detection in Nationwide Multiple Choice Tests

- In Colombia all high school students take mandatory multiple choice questions at different stages.
- There are all types of incentives for manipulations at student, school or teachers level.
- Coincidence on wrong answers maybe due to chance but only to some extent.
- Performance depends on non observables (ability, teacher quality) and others in principle observable, but in practice not observable (parents education, etc.)
- We estimate an polytomous item response model and do hypothesis testing.

Copy Detection in Nationwide Multiple Choice Tests



Proportion of couples accused of copying across exams and across samples using the ω_2^s index. The horizontal is the type-I error rate. Source: ICFES. Calculations: Authors.

Contents

State Legal Defence

- Lawsuits against the state represent over 100 billion USD (approximately 25%*ofnationalincome*). Enough to bankrupt the state.
- It is very costly to defend the state in each one of these cases.
- We used supervised and semi supervised methods to build scoring models that would be able to prioritize cases of high probability of success.
- Data set with 150.000 cases and more than 300 variables (there is space for very valuable text mining).

State Legal Defence

- Lawsuits against the state represent over 100 billion USD (approximately 25%*ofnationalincome*). Enough to bankrupt the state.
- It is very costly to defend the state in each one of these cases.
- We used supervised and semi supervised methods to build scoring models that would be able to prioritize cases of high probability of success.
- Data set with 150.000 cases and more than 300 variables (there is space for very valuable text mining).

State Legal Defence

- Lawsuits against the state represent over 100 billion USD (approximately 25%*ofnationalincome*). Enough to bankrupt the state.
- It is very costly to defend the state in each one of these cases.
- We used supervised and semi supervised methods to build scoring models that would be able to prioritize cases of high probability of success.
- Data set with 150.000 cases and more than 300 variables (there is space for very valuable text mining).

State Legal Defence

- Lawsuits against the state represent over 100 billion USD (approximately $25\% of national income$). Enough to bankrupt the state.
- It is very costly to defend the state in each one of these cases.
- We used supervised and semi supervised methods to build scoring models that would be able to prioritize cases of high probability of success.
- Data set with 150.000 cases and more than 300 variables (there is space for very valuable text mining).

- Performance

Cuadro 5: Comparación de *auc* para los tres métodos de clasificación.

Modelo	área ROC - Entrenamiento	área ROC - Validación
Logit	0.755	0.744
Boosting	0.8054	0.7626
Redes neuronales	0.7691	0.741

Contents

Tax Evasion (work in progress)

- Problem: Detect likely cases of tax evasion.
- Methodology:
  - Use information reported by firms to construct a proxy of taxes for each person.
  - Using previous years and supervised learning models, construct the best predictor of actual paid taxes.
  - Use trees to create groups of similar people.
  - Estimate the distribution within each group and detect anomalous patterns.

Tax Evasion (work in progress)

- Problem: Detect likely cases of tax evasion.
- Methodology:
  - Use information reported by firms to construct a proxy of taxes for each person.
  - Using previous years and supervised learning models, construct the best predictor of actual paid taxes.
  - Use trees to create groups of similar people.
  - Estimate the distribution within each group and detect anomalous patterns.

Contents

Money Laundry (work in progress)

- Problem: Detect anomalous transactions that may be suspicious of money laundry.
- In principle we have access to most legal data sets in Colombia: Cars transactions, properties, foreign money exchange, financial transactions, taxes, etc.
- Everything we know of unsupervised learning, scanning of data large sets, text mining, etc. is very welcome.
- This is a joint project with CESED (University of los Andes) funded by the US Embassy in Colombia.

Money Laundry (work in progress)

- Problem: Detect anomalous transactions that may be suspicious of money laundry.
- In principle we have access to most legal data sets in Colombia: Cars transactions, properties, foreign money exchange, financial transactions, taxes, etc.
- Everything we know of unsupervised learning, scanning of data large sets, text mining, etc. is very welcome.
- This is a joint project with CESED (University of los Andes) funded by the US Embassy in Colombia.

Money Laundry (work in progress)

- Problem: Detect anomalous transactions that may be suspicious of money laundry.
- In principle we have access to most legal data sets in Colombia: Cars transactions, properties, foreign money exchange, financial transactions, taxes, etc.
- Everything we know of unsupervised learning, scanning of data large sets, text mining, etc. is very welcome.
- This is a joint project with CESED (University of los Andes) funded by the US Embassy in Colombia.

Money Laundry (work in progress)

- Problem: Detect anomalous transactions that may be suspicious of money laundry.
- In principle we have access to most legal data sets in Colombia: Cars transactions, properties, foreign money exchange, financial transactions, taxes, etc.
- Everything we know of unsupervised learning, scanning of data large sets, text mining, etc. is very welcome.
- This is a joint project with CESED (University of los Andes) funded by the US Embassy in Colombia.

Contents

Other Projects

- Marketing (Young and Rubican): Social newtworks
- Call Center (Millenium): Speech transcription, email duplications, community management, etc.

Other Projects

- Marketing (Young and Rubican): Social newtworks
- Call Center (Millenium): Speech transcription, email duplications, community management, etc.

Contents

Who are we?

- Quantil (www.quantil.com.co) it a company dedicated to the application of mathematics to real world problems.
- Our main philosophical guidelines are bringing scientific standards to analyse and deliver any customers demand for services.
- The company was founded in 2008 by two PhDs in mathematics.

Who are we?

- Quantil (www.quantil.com.co) it a company dedicated to the application of mathematics to real world problems.
- Our main philosophical guidelines are bringing scientific standards to analyse and deliver any customers demand for services.
- The company was founded in 2008 by two PhDs in mathematics.

Who are we?

- Quantil (www.quantil.com.co) it a company dedicated to the application of mathematics to real world problems.
- Our main philosophical guidelines are bringing scientific standards to analyse and deliver any customers demand for services.
- The company was founded in 2008 by two PhDs in mathematics.

Who are we?

- Quantil (www.quantil.com.co) it a company dedicated to the application of mathematics to real world problems.
- Our main philosophical guidelines are bringing scientific standards to analyse and deliver any customers demand for services.
- The company was founded in 2008 by two PhDs in mathematics.

Who are we?

- We have three main business areas and a research and development unit:
  1. Economic modelling: Energy, health and communications sectors.
  2. Mathematical finance: hedging, valuation, quantitative risk and assets management, algorithmic trading, etc.
  3. Data mining: traditional databases, social networks.

Who are we?

- We have worked with approximately 80 organizations including:
  - Ministries (health, finance)
  - Government agencies (national planing department, energy, and communications regulatory agencies)
  - Many private companies (financial institutions, marketing agencies, call centers, commodities producers, etc.)
  - Multilaterals (World Bank, Interamerican Development Bank).

Who are we?

- Some of our our colleagues are now studying abroad: UCSD, Columbia, Carnegie Melon.

Who are we?

Thank you very much! alvaro.riascos@quantil.com.co