Mining Government Data-sets for Detecting Suspicious Behavior

Alvaro J. Riascos Villegas
University of los Andes and Quantil

Data Mining for Corruption Hunters
3rd Biennial ICHA Meeting

December 9, 2014

Contents

Introduction

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

Introduction

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

Introduction

- This short presentation provides a few examples of successful mining of Government 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.
- Payments are determined using information provided by HMOs.
- Potential problems: Risk of manipulation, misreport, errors, etc.
- Task: Design a system that helps the Ministry to detect all sorts of anomalous behavior.

Health Records

- Colombia has a mandatory competitive health insurance system.
- Payments are determined using information provided by HMOs.
- Potential problems: Risk of manipulation, misreport, errors, etc.
- Task: Design a system that helps the Ministry to detect all sorts of anomalous behavior.

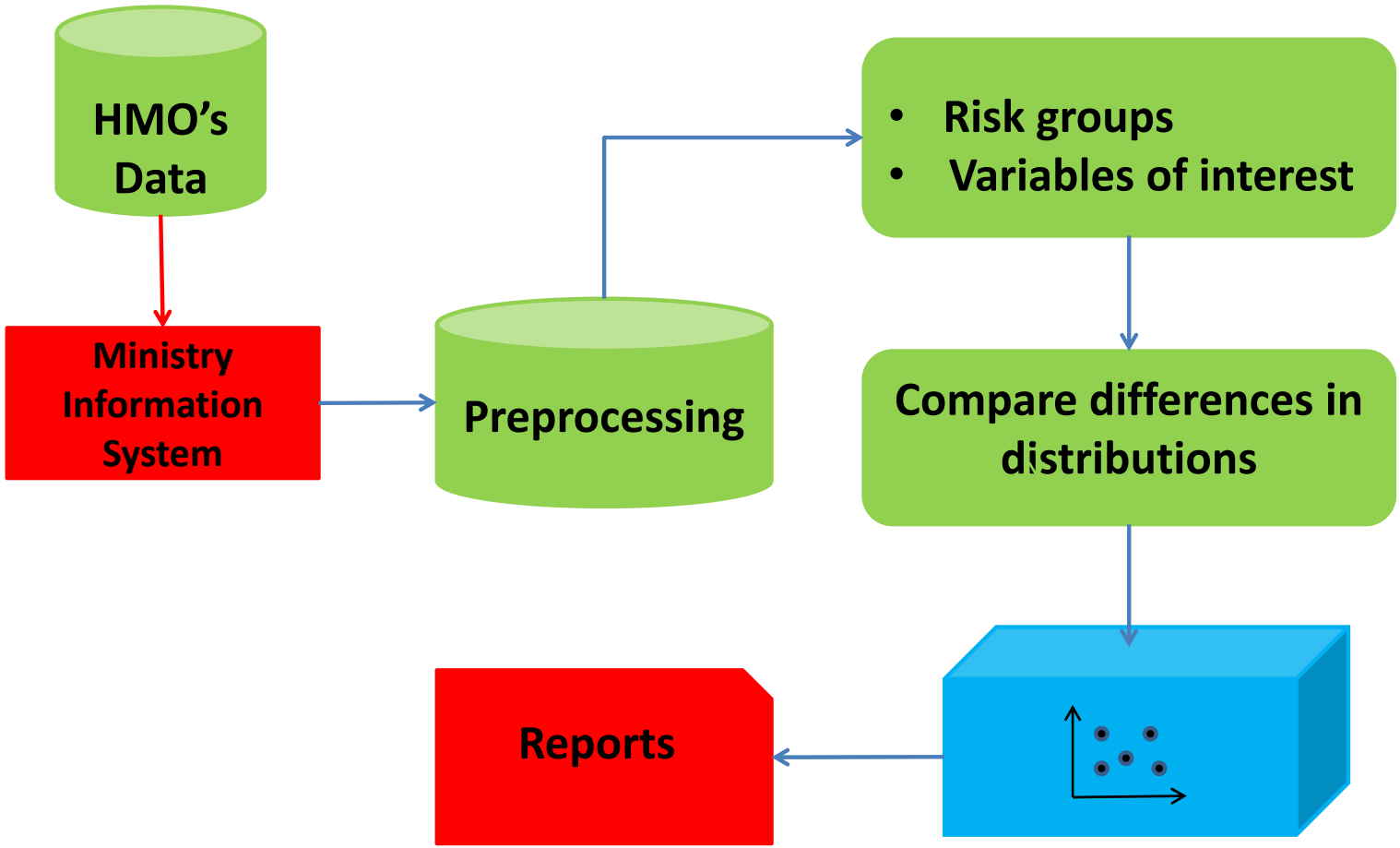
Health Records

- Colombia has a mandatory competitive health insurance system.
- Payments are determined using information provided by HMOs.
- Potential problems: Risk of manipulation, misreport, errors, etc.
- Task: Design a system that helps the Ministry to detect all sorts of anomalous behavior.

Health Records

- Colombia has a mandatory competitive health insurance system.
- Payments are determined using information provided by HMOs.
- Potential problems: Risk of manipulation, misreport, errors, etc.
- Task: Design a system that helps the Ministry to detect all sorts of anomalous behavior.

Health Records



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.
  - Estimate the distribution within homogeneous groups 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.
  - Estimate the distribution within homogeneous groups 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.
  - Estimate the distribution within homogeneous groups and detect anomalous patterns.

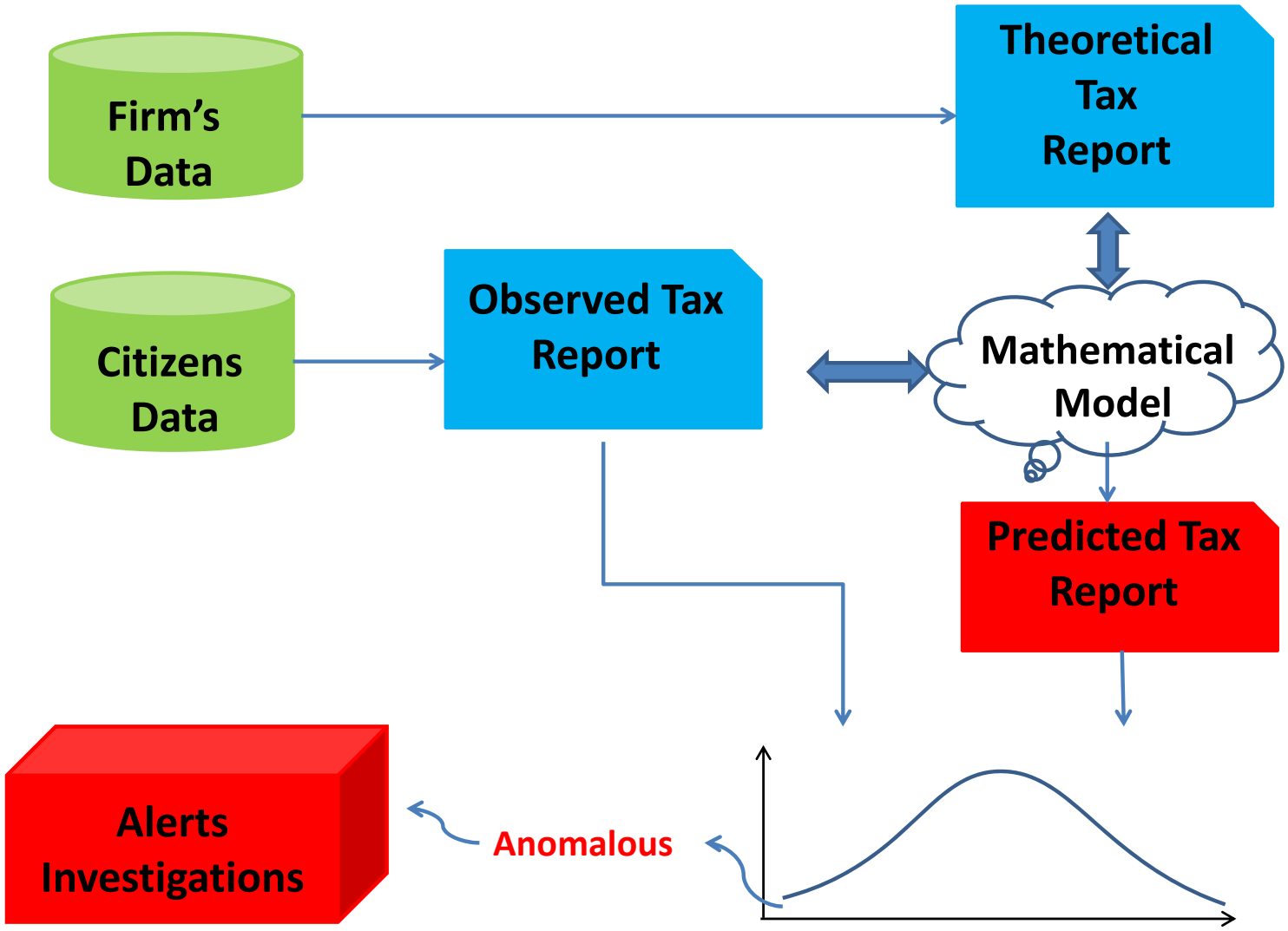
Tax Evasion (work in progress)

- 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.
  - Estimate the distribution within homogeneous groups 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.
  - Estimate the distribution within homogeneous groups and detect anomalous patterns.

Tax Evasion


Firm's Data
Theoretical Tax Report
Citizens Data
Observed Tax Report
Mathematical Model
Predicted Tax Report
Alerts Investigations
Anomalous

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.
- Methodology: Everything we know of unsupervised learning, scanning of data large sets, text mining, etc. is very welcome.
- 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.
- Methodology: Everything we know of unsupervised learning, scanning of data large sets, text mining, etc. is very welcome.
- 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.
- Methodology: Everything we know of unsupervised learning, scanning of data large sets, text mining, etc. is very welcome.
- 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.
- Methodology: Everything we know of unsupervised learning, scanning of data large sets, text mining, etc. is very welcome.
- Joint project with CESED (University of los Andes) funded by the US Embassy in Colombia.

Contents

Data Science Framework

- How it works? You need:
  1. Data, storage capabilities and computational power.
  2. Expert advice to inform research.
  3. Data scientist to run machine and/or statistical learning algorithms.
- This framework complements traditional investigation techniques:
  1. Generates alerts.
  2. Informs on where and what to investigate.
  3. Allows for the automated and frequent revision of large datasets.

Data Science Framework

- How it works? You need:
  1. Data, storage capabilities and computational power.
  2. Expert advice to inform research.
  3. Data scientist to run machine and/or statistical learning algorithms.
- This framework complements traditional investigation techniques:
  1. Generates alerts.
  2. Informs on where and what to investigate.
  3. Allows for the automated and frequent revision of large datasets.

Data Science Framework

- How it works? You need:
  1. Data, storage capabilities and computational power.
  2. Expert advice to inform research.
  3. Data scientist to run machine and/or statistical learning algorithms.
- This framework complements traditional investigation techniques:
  1. Generates alerts.
  2. Informs on where and what to investigate.
  3. Allows for the automated and frequent revision of large datasets.

Data Science Framework

- How it works? You need:
  1. Data, storage capabilities and computational power.
  2. Expert advice to inform research.
  3. Data scientist to run machine and/or statistical learning algorithms.
- This framework complements traditional investigation techniques:
  1. Generates alerts.
  2. Informs on where and what to investigate.
  3. Allows for the automated and frequent revision of large datasets.

Data Science Framework

- How it works? You need:
  1. Data, storage capabilities and computational power.
  2. Expert advice to inform research.
  3. Data scientist to run machine and/or statistical learning algorithms.
- This framework complements traditional investigation techniques:
  1. Generates alerts.
  2. Informs on where and what to investigate.
  3. Allows for the automated and frequent revision of large datasets.

Data Science Framework

- How it works? You need:
  1. Data, storage capabilities and computational power.
  2. Expert advice to inform research.
  3. Data scientist to run machine and/or statistical learning algorithms.
- This framework complements traditional investigation techniques:
  1. Generates alerts.
  2. Informs on where and what to investigate.
  3. Allows for the automated and frequent revision of large datasets.

Data Science Framework

- How it works? You need:
  1. Data, storage capabilities and computational power.
  2. Expert advice to inform research.
  3. Data scientist to run machine and/or statistical learning algorithms.
- This framework complements traditional investigation techniques:
  1. Generates alerts.
  2. Informs on where and what to investigate.
  3. Allows for the automated and frequent revision of large datasets.

Data Science Framework

- How it works? You need:
  1. Data, storage capabilities and computational power.
  2. Expert advice to inform research.
  3. Data scientist to run machine and/or statistical learning algorithms.
- This framework complements traditional investigation techniques:
  1. Generates alerts.
  2. Informs on where and what to investigate.
  3. Allows for the automated and frequent revision of large datasets.

Thank you very much!

alvaro.riascos@quantil.com.co