A paradigm shift from Surveys to Big Data in Financial Market Research

A. Chinomona

Universitat Pompeu Fabra (UPF), Barcelona, Spain

26 October 2018
Outline

- Introduction
- Survey Data Analysis
- Big Data Analytics
- Big Data Manipulation Tools
- Application: Big Data vs Survey Data
- Conclusion
- References
Motivation

“Surveys are dead! We’re now living in the “Big Data” era—a world of voluminous, high velocity, and increasingly varied data sources. Surveys have been the “work-horse” of market research for nearly a century, but long lead times, small sample sizes, declining participation, and rising costs are making it far more difficult to conduct good surveys today than in the past”. Michael Link (President of Abt SRBI one of the US’s survey, opinion, and policy research organizations)

Availability of sophisticated computers and availability of data: fast, continuous/real time, structured/unstructured, complex and variable (ever changing) has made it easy to collect and store enormous and complex datasets termed Big Data as proposed by Diebold (2003)
### Introduction

**Where are we? From Surveys to Big Data**

#### Source of Data
- Internet data
- Social media
- Website metadata e.g. searches, adverts, transactions etc
- The Internet of Things (IoT)
- Retail transaction data
- Administrative data
- Commerically available databases

#### Format
- Structured and unstructured, semi-structured, quasi-structured
text, images, videos, file formats e.g. parquet

#### Storage and Retrieval
- Databases, Rel, databases
- Cloud memory, e.g. icloud,
Amazon Big Data
- Google cloud, Drop box,
Data warehouses
- Hadoop sequence files, contextual meta data

#### Processing
- Big Data Analytics
  - Batch, real time, machine learning, neural networks
  - MapReduce and Hadoop

---

**Data-based Research**

**Survey**
- Traditional questionnaire based surveys
- Examples
  - DHS, NHIS, the Income and Exp Surveys

**Survey Data Analysis**
- Typically supported by Statistical theory, main focus is on summarization, statistical inference and prediction

---
Survey Data Analysis

- Analysis based on **flat rectangular data**.
- Analysis uses **conventional statistical techniques** supported by the fundamental theory of sampling, probability and statistical inference to explain the stochastic processes underlying the dynamics of the phenomena under investigation.
- Data are usually collected using such tools as questionnaires.
Big Data Analytics

**Definition and Characteristics**

- **Volume**: Vast amounts of data generated through large-scale digitalization of information.
- **Velocity**: Speed at which data are generated.
- **Variety**: Different types and forms of data, e.g., structured and unstructured.
- **Veracity**: Level of quality, accuracy, and uncertainty in data and their sources.
- **Value**: Socio-economic contribution of big data.
Big Data Analytics

Big Data Manipulation Tools

- The sheer size and complexities necessitate special tools for extracting and analyzing Big Data e.g SQL and NoSQL.
- Most operations are run on a cluster of computers provided by such providers as Amazon, Google etc using techniques such as MapReduce and Hadoop.
Application: Big Data vs Survey Data

Stocks Data

- I used data from the Johannesburg Stock Exchange (JSE) for 2017.
- The data comprise stock data i.e. prices, volume, dividend yield collected in real time and ratio between the current share price and the expected earnings on the share.
- Used whole dataset (real time stocks prices) as Big Data.
  - the data are stored in specialized Time Series Database (TSDB) (relational databases) based on open source NoSQL on the Rhodes University server.
  - I used the PostgreSQL RHadoop to retrieve the data.
- I then used a complex survey design to draw out a sample
  - Stratified by Sectors: SA Resources, SA Financials and SA Industrials.
  - PSU weeks
- I used both Big Data Analytics and Survey Data Analysis techniques and compare the results.
Computing

pgAdmin 4 and PostgreSQL:

- pgAdmin 4 is an open source management tool for PostgreSQL (an object relational database management system)
- PostgreSQL:
  - a project designed to use different programming languages such as C/C++, Java, Python and Open Database Connectivity (ODBC) and supports text, images, sounds etc.
  - supports the SQL standard including features such as complex SQL queries.
  - It has several functions to manage a database
Computing Databases with R

Connecting to the PostgreSQL database in R

- I using RPostgreSQL package
- There are six settings needed to make a connection
  - **Driver** = Postgres SQL driver, **Server** = network path to the database,
  - **Database** = the name of the schema, **UID** = the user's network ID or server local account,
  - **PWD** = the account’s password, **Port** = 5432

My R code

```r
> pw <- "password"
> con <- dbConnect(RPostgres::Postgres(),
    host="cs202.ict.ru.ac.za",
    port=5432, dbname="amos", user="amos",
    password=pw)
```
I drew sample using a complex sampling design (stratified cluster sampling design) with strata (Sectors) and clustered with \( PSU = \) weeks using R.

Plots of the Data

- The Time Series plots

**Big Data**

**Weekly Average Ratio**

**Survey Data**

**Weekly Average Dividend Yield**
Plots of the Data

Survey Data

Weekly Average Actual Closing Price

Big Data

Weekly Average Actual Closing Price

Weekly Average Actual Volume

Weekly Average Actual Volume
Results

Summaries

> complexdesign<-svydesign(id=~week,
  strata=~sector,data=sample, nest=TRUE)
> svymean(~div_yield,complexdesign,deff=TRUE)

<table>
<thead>
<tr>
<th>Mean</th>
<th>Suvery Data Analysis</th>
<th>Big Data Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Closing</td>
<td>2092.6360 (15.562)</td>
<td>2123.1300 (28.1300)</td>
</tr>
<tr>
<td>Dividend Yield</td>
<td>3.0699 (0.0249)</td>
<td>3.0305 (0.0262)</td>
</tr>
<tr>
<td>Per Ratio</td>
<td>21.8355 (0.5194)</td>
<td>22.9002 (0.5852)</td>
</tr>
</tbody>
</table>
Results

Time Series Analysis

- Note a typical time series, as developed by Box and Jenkins (1976) is explained by Autoregressive (AR), Moving Average (MA) and integrated terms, Thus A time series $X_t$ is said to be ARIMA of order $(p, d, q)$ given by

$$Y_t = \mu + \sum_{i=1}^{p} \alpha_i Y_{t-1} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} + \varepsilon_t$$

where $Y_t = \Delta^d X_t$ is the differenced series to achieve stationarity

- I fitted an ARIMA model for the dividend yield series for both survey and Big Data.

- The `auto.arima` function in R runs several combinations of models and selects the most parsimonious model was used.

- Big Data Time Series analysis falls into the supervised learning prediction framework
Results

Time Series Analysis

Hence an $ARIMA(1, 0, 4)$ was fitted for the Big Data and an $ARIMA(5, 1, 0)$ for surveys data for the dividend yield series

### Table 1: Estimates of $ARIMA(1, 0, 4)$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std Err</th>
<th>$p − \text{value}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>22.9001</td>
<td>2.8361</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.9440</td>
<td>0.0038</td>
<td>0.0087</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.0237</td>
<td>0.0080</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.0048</td>
<td>0.0077</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-0.4744</td>
<td>0.0073</td>
<td>0.0016</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.0116</td>
<td>0.0075</td>
<td>0.0030</td>
</tr>
</tbody>
</table>

### Table 2: Estimates of $ARIMA(5, 1, 0)$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std Err</th>
<th>$p − \text{value}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>-0.8422</td>
<td>0.0112</td>
<td>0.005</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>-0.6761</td>
<td>0.0142</td>
<td>0.001</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>-0.517</td>
<td>0.015</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>-0.3300</td>
<td>0.0142</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td>-0.1601</td>
<td>0.0112</td>
<td>0.009</td>
</tr>
</tbody>
</table>

The resulting models are:

**Big Data:**

$$\hat{X}_t = 22.9001 + 0.944X_{t−1} + 0.0237\varepsilon_{t−1} − 0.0048\varepsilon_{t−2} − 0.4744\varepsilon_{t−3} + 0.0116\varepsilon_{t−4}$$

**Survey Data:**

$$Y_t = X_t − X_{t−1} = −0.8422X_{t−1} − 0.6761X_{t−2} − 0.517X_{t−3} − 0.33X_{t−4} − 0.1601X_{t−5}$$

$$\therefore \hat{X}_t = 0.1578X_{t−1} − 0.6716X_{t−2} − 0.517X_{t−3} − 0.33X_{t−4} − 0.1601X_{t−5}$$
Big Data

Forecasting

Survey
Conclusion

- There is no much difference in the results.
References


A Big (data) Thank You