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

A. Chinomona

Universitat Pompeu Fabra (UPF), Barcelona, Spain

26 October 2018

RHODES UNIVERSITY

Grahamstown • 6140 • South Africa

Outline

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

Introduction

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

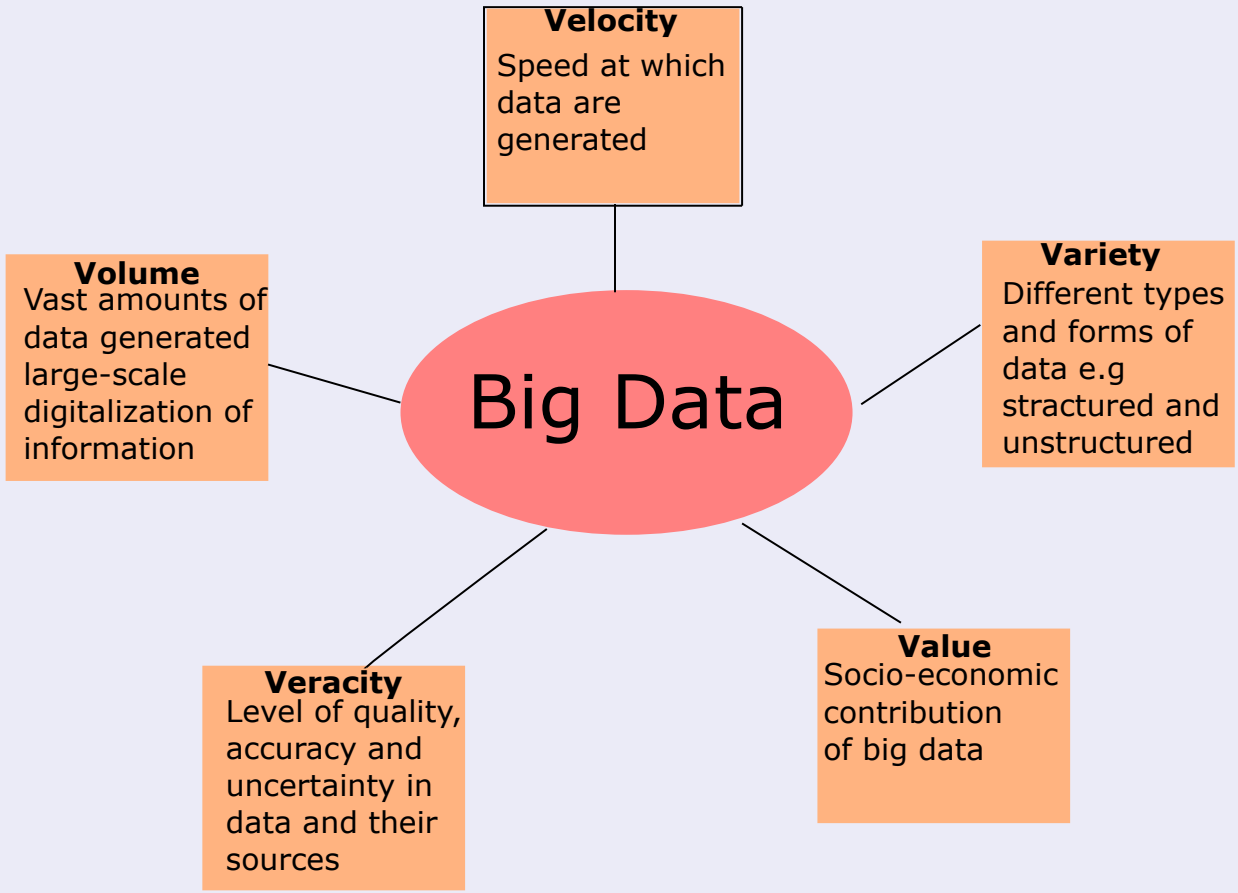
		Source of Data	Format	Storage and Retrieval	Processing
Data-based Research	**Big Data**	Internet data Social media Website metadata e.g searches, adverts, transac-tions etc The Internet of Things (IoT) Retail transaction data Administrative data Commercally available databases	Structured and unstructured, semi-structred. quasi-structured text, images, videos, file formats e.g parquet	Databases, Rel, databases Cloud memory, e.g icloud, Amazon Big Data Google cloud, Drop box, Data warehouses Hadoop sequence files, contextual meta data Data mining tech SQL and NoSQL	Big Data Analytics Batch, real time, machine learning, nueral networks MapReduce and Hadoop
	Survey	Traditional questionnaire based surveys Examples DHS, NHIS, the Income and Exp Surveys	Flat, Rectangular, Structured typically, in a spreadsheet	Usual computer storage and external hard drives, No need for programming language	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



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:

- pgAmin 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

```
>pw<-"password"
>con<-dbConnect(RPostgres::Postgres(),
        host="cs202.ict.ru.ac.za",
                port=5432,dbname="amos",user="amos",
                    password=pw)
```

Computing

Survey in R

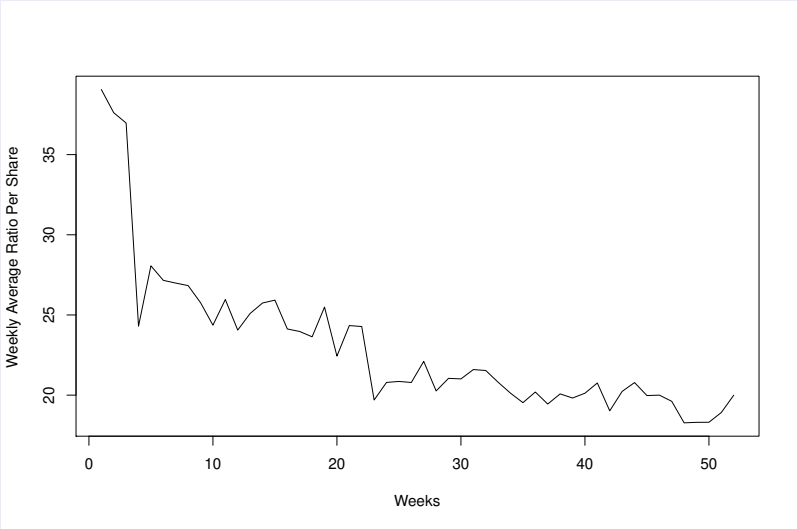
- I drew sample using a complex sampling design (stratified cluster sampling design) with strata (Sectors) and clustered with $PSU =$ weeks using R.
- I used techniques developed by Lumley (2010) for analysis of survey data using theory in Cochran (1977) and Lohr (2010).

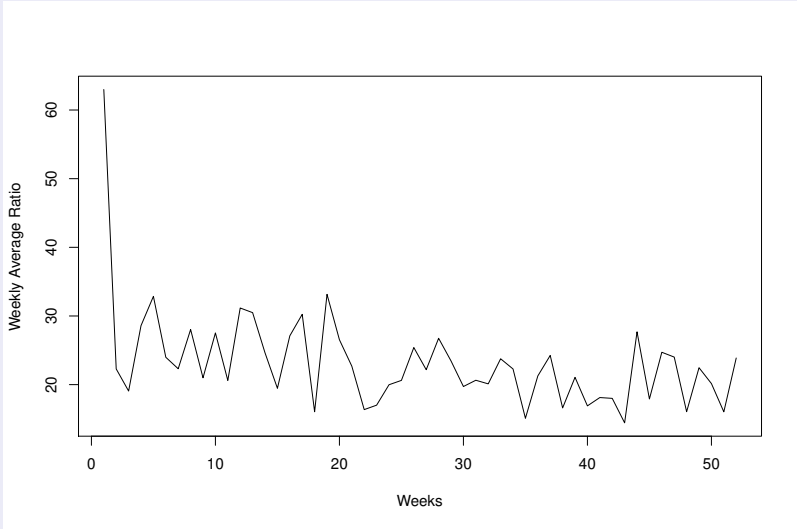
Plots of the Data

- The Time Sieres plots

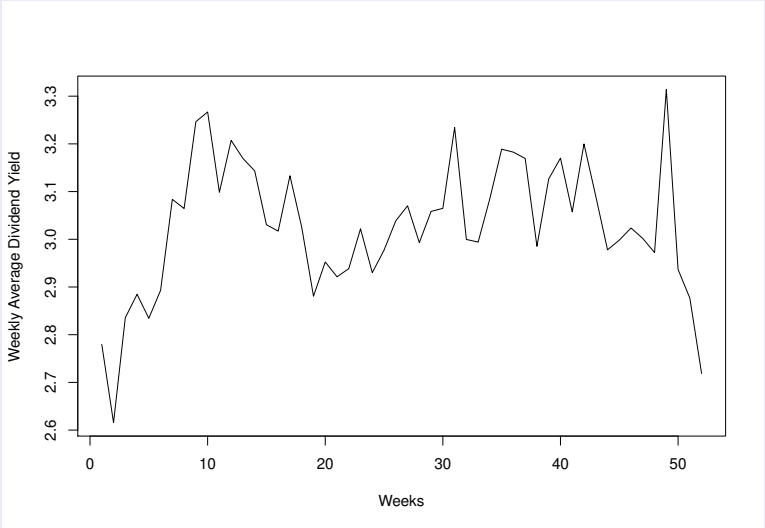
Big Data | **Survey Data**

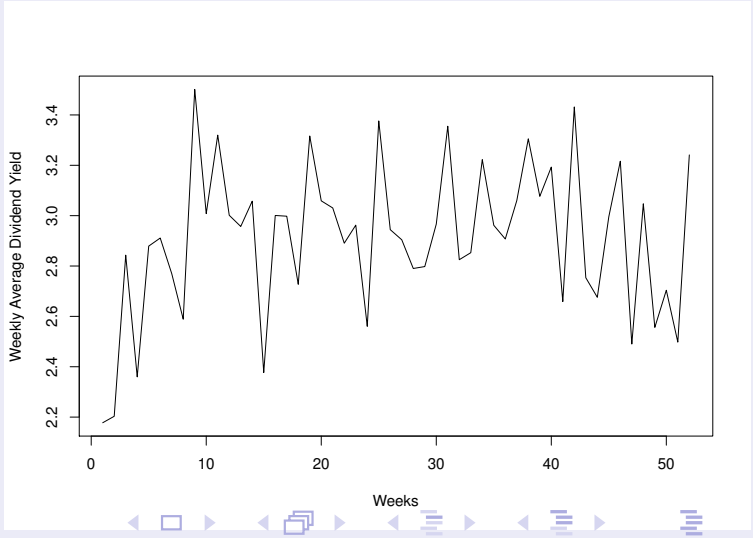
Weekly Average Ratio





Weekly Average Dividend Yield



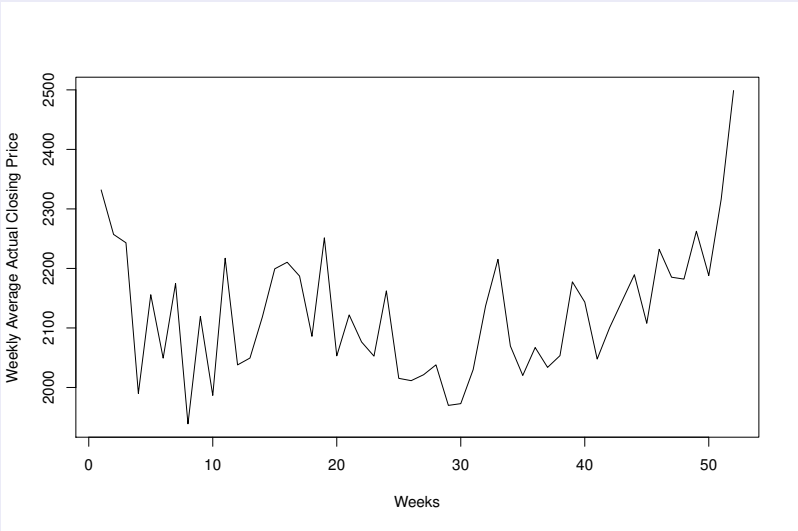


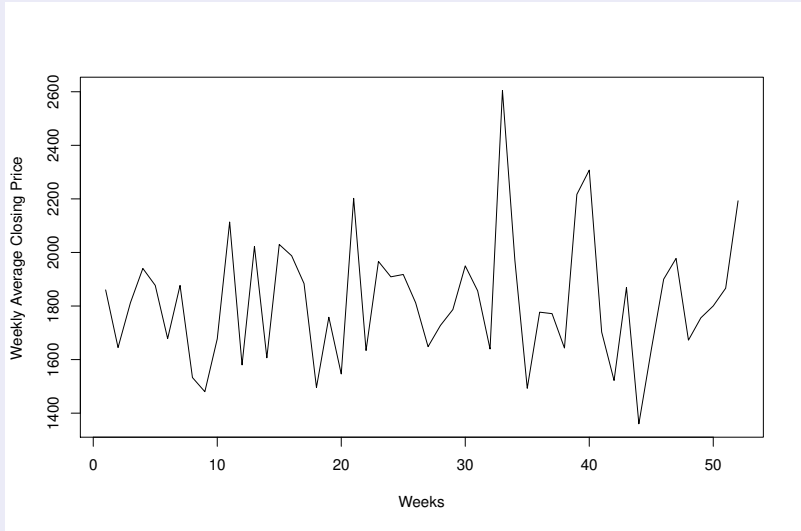
Plots of the Data

Survey Data

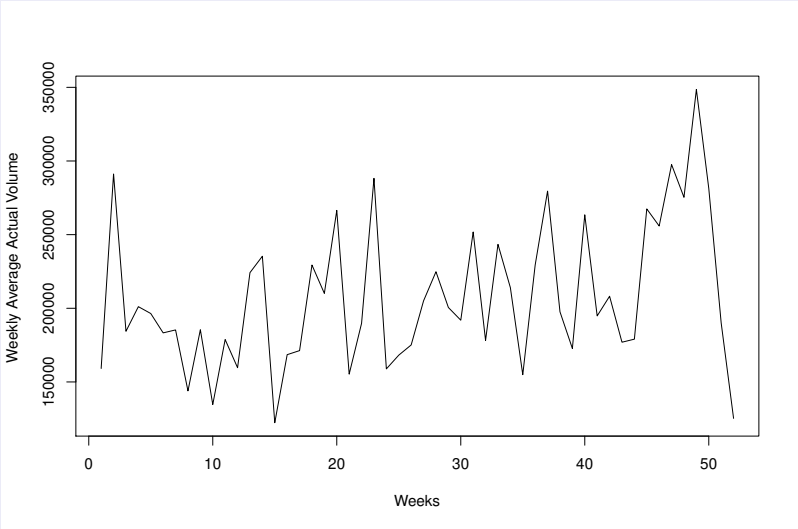
Big Data

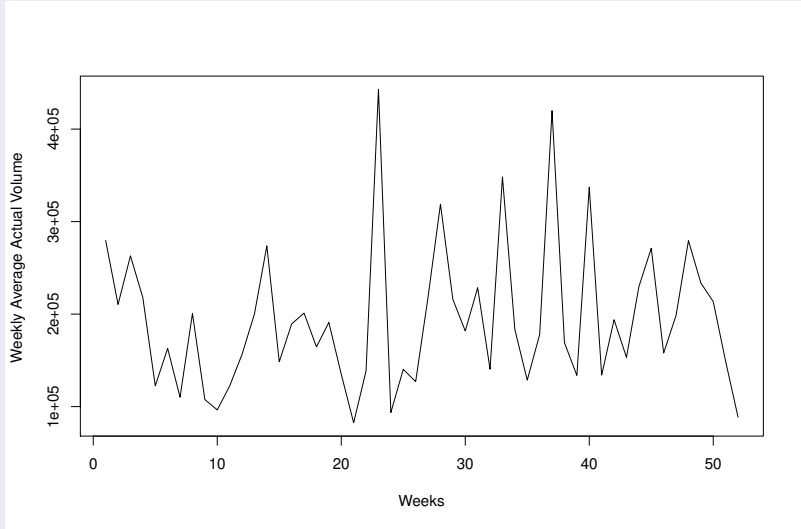
Weekly Average Actual Closing Price





Weekly Average Actual Volume





Results

Summaries

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

Mean	Suvery Data Analysis	Big Data Analytics
Actual Closing	2092.6360 (15.562)	2123.1300 (28.1300)
Dividend Yield	3.0699 (0.0249)	3.0305 (0.0262)
Per Ratio	21.8355 (0.5194)	22.9002 (0.5852)

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

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

  where $Y_t = \Delta^d X_t$ is the differenced series to achiecve 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 $ARIMA(1,0,4)$

Parameter	Estimate	Std Err	$p-$value
μ	22.9001	2.8361	< 0.001
α_1	0.9440	0.0038	0.0087
β_1	0.0237	0.0080	< 0.001
β_2	-0.0048	0.0077	< 0.001
β_3	-0.4744	0.0073	0.0016
β_4	0.0116	0.0075	0.0030

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

Parameter	Estimate	Std Err	$p-$value
α_1	-0.8422	0.0112	0.005
α_2	-0.6761	0.0142	0.001
α_3	-0.517	0.015	< 0.001
α_4	-0.3300	0.0142	< 0.001
α_5	-0.1601	0.0112	0.009

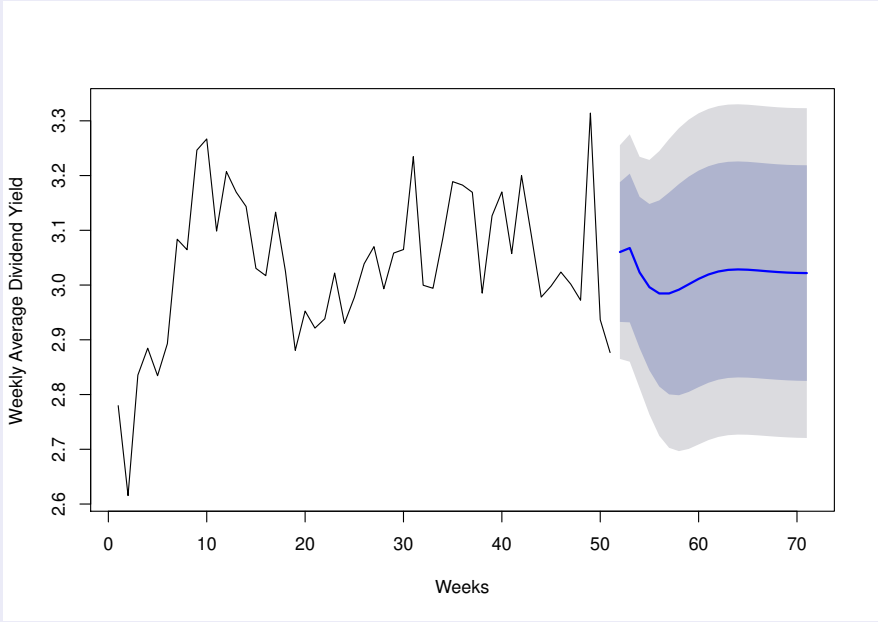
The resulting models are:

Big Data: $$\widehat{X_t} = 22.9001 + 0.944X_{t-1} + 0.0237\varepsilon_{t-1} - 0.0048\varepsilon_{t-2} - 0.4744\varepsilon_{t-3} + 0.0116\varepsilon_{t-4}$$

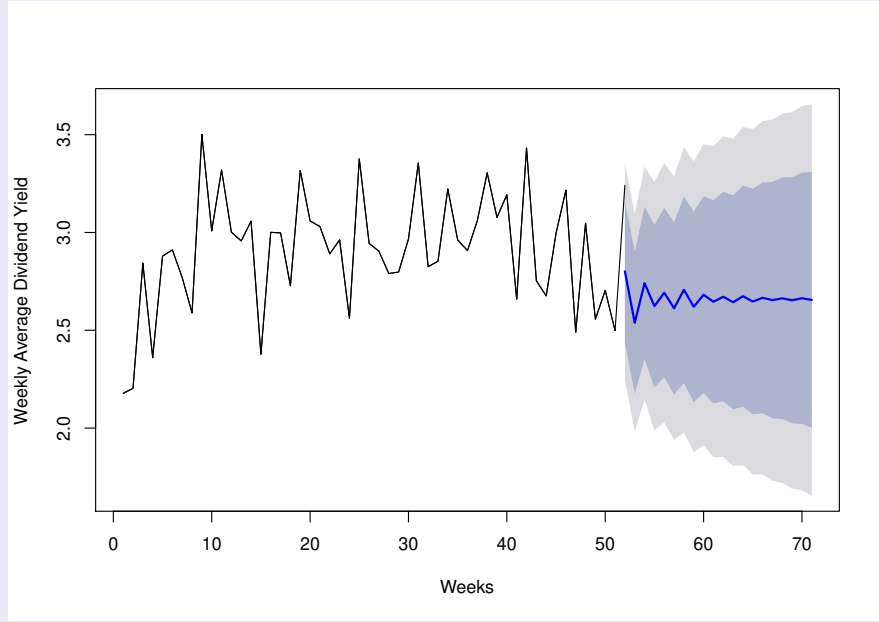
Survey Data:
$$\begin{aligned} Y_t &= X_t - X_{t-1} \\ &= -.8422X_{t-1} - 0.6716X_{t-2} - 0.517X_{t-3} - 0.33X_{t-4} - 0.1601X_{t-5} \\ \therefore \widehat{X_t} &= 0.1578X_{t-1} - 0.6716X_{t-2} - 0.517X_{t-3} - 0.33X_{t-4} - 0.1601X_{t-5} \end{aligned}$$

Forecasting

Big Data



Survey



Conclusion

- There is no much difference in the results.

References

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*.

Cochran, W. G. (1977). *Sampling Techniques*. Wiley Series in Probability and Mathematical Statistics.

Diebold, F. X. (2003). Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting. *Advances in Economics and Econometrics: Theory and Applications, Eight World Congress of the Econometrics Society*, pages Cambridge University Press 115–122.

Lohr, S. L. (2010). *Sampling: Design and Analysis, 2nd Edition.* Cengage Learning.

Lumley, T. (2010). *Complex Surveys: A guide to Analysis Using R.* John Wiley and Sons Inc., Washington, USA.


A Big (data) Thank You