Data sources for MGT285

May 22, 2013

Hugh Crockford

Contents

1 Introduction

There is a large amount of data available from a variety of credible and not so credible sources online. Historical time series of macroeconomic indicators and indexes can be useful variables to consider including in any long duration time series model, and the rise of the 'social web' has produced reams of data to aid and inform decision making.

This vignette describes a variety of data sources that may be useful during any modelling project.

2 Sources of Historical Time Series data

Variables such as population, income, and CPI are economy wide metrics that can explain some of the variability present in a dataset. Historical pricing information for commodities and financial instruments are also potentially important when developing costing/pricing or demand models. Many exchange sites will offer these data for a hefty subscription fee, however most can be found online for free

2.1 Financial Time Series - Index and pricing data

Some data can be found with a simple Google search, however many sites are of questionable quality or will request payment for full access.

Ycharts [1] appears to be the best bang-for-buck at $49 month. This site has many data series for plotting and download, and also includes an Excel add-in with purchase of professional membership. Approximate data can also be extracted from freely available graphs as described in Appendix C

A Bloomberg terminal subscription provides access to a plethora of Industry grade datasets, including real-time market data and historical quotes. A subscription to Bloomberg is a minimum $2000 pcm however many financial business' have a subscription that can be used to download series as needed. The UC Library system unfortunately does not have access to Bloomberg data. Bloomberg also offers an API [2] that allows data to be dynamically loaded into excel.

Quandl [3] appears to be the most useful online data source, with large amounts of quality data freely accessible. It aggregates over 5 million financial, economic, and social datasets from around the web with primary reference information. All datasets are manually accessible and downloadable from the search engine on their homepage, and can be accessed from within R using the package 'Quandl', and dynamically added to a spreadsheet via an Excel plugin.

2.2 Demographic information

Population and demographic information can be found around the web, however the most trusted source is American Census Bureau. Their website is a little difficult to navigate, and the 'American fact-finder'[4] is the search engine to locate area specific data. The most powerful way to access the census data is through their API, with which quereies can be built and run over multiple geographic areas and variables.

[1] http://ycharts.com/
[2] http://www.openbloomberg.com/open-api/
[3] http://www.quandl.com/
[4] http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml

3 Social Media Data

The explosion in social media has allowed the monitoring of massive amounts of user generated data to generate insights into product choices, pending pandemics, and even stock prices.

3.1 Google Trends/Search volumes

Most consumers employ a search engine for everything from what car to buy to finding out if they have a cold or tuberculosis. Tracking search volume across time can provide insights into what the 'crowd' is researching, from which various inferences can be made. The first application of this technology was when Google developed algorithms to predict Flu epidemics based on search volumes, diagnosing epidemics much sooner than the mighty CDC [1]. Google Trends [5] is a demonstration of this capability available for everyone to use. Keywords can be compared (e.g. a product and it's competitor) and relative search volumes found, with news items associated with peaks in search volumes. The data can be downloaded into a csv file (comma separated value) that is easily loaded into excel, and queries can be based on location, time frame and category. The Google Trends API can also be accessed by R packages 'rGtrend' and 'RGoogleTrends'

Related to Google trends is Wikipedia page views, which has shown to be indicative of future stock price moves [2]. Wikipedia page statistics can be accessed manually [6], or programatically via a RESTful API [3].

3.2 Twitter

Microblogging platforms such as Twitter generate over 400 million data points each day, data that can be analysed to reveal people's brand awareness and perceptions, and can reveal future market events [4]. Many people are actively watching social media and trading on this data, with twitter sentiment analysis shown to beat various metrics in predicting DJIA closing values [5]. The importance traders and their algorithmic trading programs place on information from social media was demonstrated in May 2013 when the Associated press' twitter account was hacked and a false tweet reporting the White House had been bombed was released. The false tweet erased $200 billion from the US Stock Market in 2 minutes, with the Dow falling almost 150 points and oil and tbond futures following [6].

Historical tweets (last 7 days) can be accessed from Twitter's API [7] and interpreted using Natural Language Processing tools. There are also various services that present trending tweets geographically [8] and by hashtag [9]

[5] http://www.google.com/trends/explore#cmpt=q
[6] http://stats.grok.se/
[7] https://dev.twitter.com/docs/api
[8] https://dev.twitter.com/docs/api
[9] http://www.hashtags.org/

Appendices

A R Resources

R is an open source statistical programming language widely used in academia and business. While the initial learning curve can be steep, the scripting and scaling capabilities mean an initial time investment will pay dividends if any serious modelling is being completed. These scripting capabilities also force analysts to explicitly state assumptions, constants and equations, allowing easier oversight and validation when compared to an excel spreadsheet model. Errors in complicated spreadsheets have been implicated in some high profile cases recently (London whale, MF global, Lehman [7]) with incorrect cell references or equation errors proving disastrous for all involved. In addition, many of the techniques discussed in this vignette are implemented directly in R, allowing seamless integration of data collection and modelling.

There are many books available to assist learning R, and a wealth of online information. Coursera [10] offers a free online course in 'Computing for Data Analysis', and other statistics courses also use R. For those not used to the command line interface, R Studio [11] is a free GUI (graphical user interface) which has many tools to assist learning R.

There are numerous classes on main campus available to learn R, ranging from a basic introduction given in most applied stat classes to more advanced classes that require a solid grasp of the program.

B Web Application Programming Interface's (API)

Many data rich websites recognise navigating menu's and downloading individual files is laborious so have developed web API's to allow scripted access to data. For any project requiring many queries of a database, it is worth scripting the requests (commonly HTTP GET/POST) to allow many iteration's to be run across a range of variables.

The predominant web API is REST (Representational state transfer), which process HTTP requests and commonly returns data as xml or JSON. Each REST server will have its own methods and documentation will be provided to allow scripted query development.

Some api's (e.g. quandl,twitter) will require a registration key so that usage can be tracked.

C Extracting data from a Plot

Often a plot can be found with required time series, however the data used to generate the figure are unavailable. An approximation of underlying data can be generated by reading the position of lines/points versus calibrated known points. Numerous tools exist to complete this task, the easiest to use being WebPlotDigitizer, an online tool [12]. A graph is loaded, calibration points selected and their corresponding values entered, then color thresholds set for lines/points. The software then automatically selects points along the line, and the resulting data can be downloaded as a csv. This task can also be completed within R using the 'digitize' package.

[10] https://www.coursera.org/
[11] http://www.rstudio.com/
[12] http://arohatgi.info/WebPlotDigitizer/

D Bibliography

References

[1] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–4, February 2009.

[2] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y Kenett, H Eugene Stanley, and Tobias Preis. Quantifying wikipedia usage patterns before stock market moves. *Scientific reports*, 3:1801, May 2013.

[3] M-H. Peetz, E. Meij, and M. de Rijke. OpenGeist: Insight in the stream of page views on Wikipedia. In *SIGIR 2012 Workshop on Time-aware Information Access*, 2012.

[4] Eduardo J. Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Correlating financial time series with micro-blogging activity. *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, page 513, 2012.

[5] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, March 2011.

[6] Twitter Hoax Sparks Swift Stock Swoon - WSJ.com. http://online.wsj.com/article/SB10001424127887323735604 2013.

[7] How Excel is ruining the world - The Term Sheet: Fortune's deals blogTerm Sheet. http://finance.fortune.cnn.com/2013/04/17/rogoff-reinhart-excel-errors/, 2013.