SMU

SINGAPORE MANAGEMENT
UNIVERSITY

DSA 301

Term Project Options, 2024

Table of Contents

Overview

Each group can select one topic to develop for their project proposal, due in week 4.

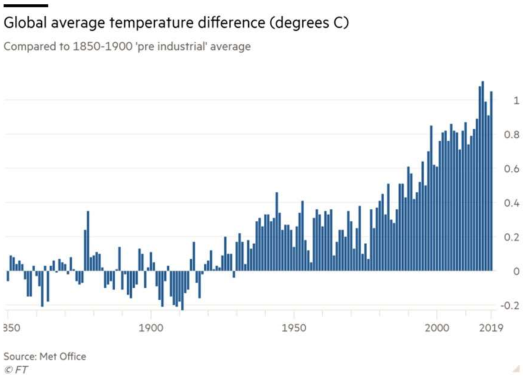
Proposals outside of above list can also be entertained. In this case, email a short (1 paragraph) informal summary and discussing approximate data sources and analysis outline by end of week 3 so I can consider viability prior to lengthier proposal in week 4.

Option 1: Climate Change, Human-caused or not?

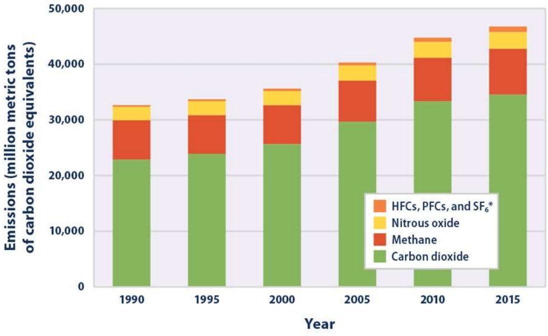
Long term climate change research has been modelled as an interaction between the following 3 variables:

1. "Green-house" gas emissions
2. Global average temperatures
3. Sea levels

Other indicators (either causal factors or outcomes) have been proposed: deforestation, droughts and species extinction. Long term economic costs of the forgoing will need to be balanced against the short term price of reducing emissions: carbon removal, higher cost of substitutes for selected gasses, and higher short term costs of renewable energy.



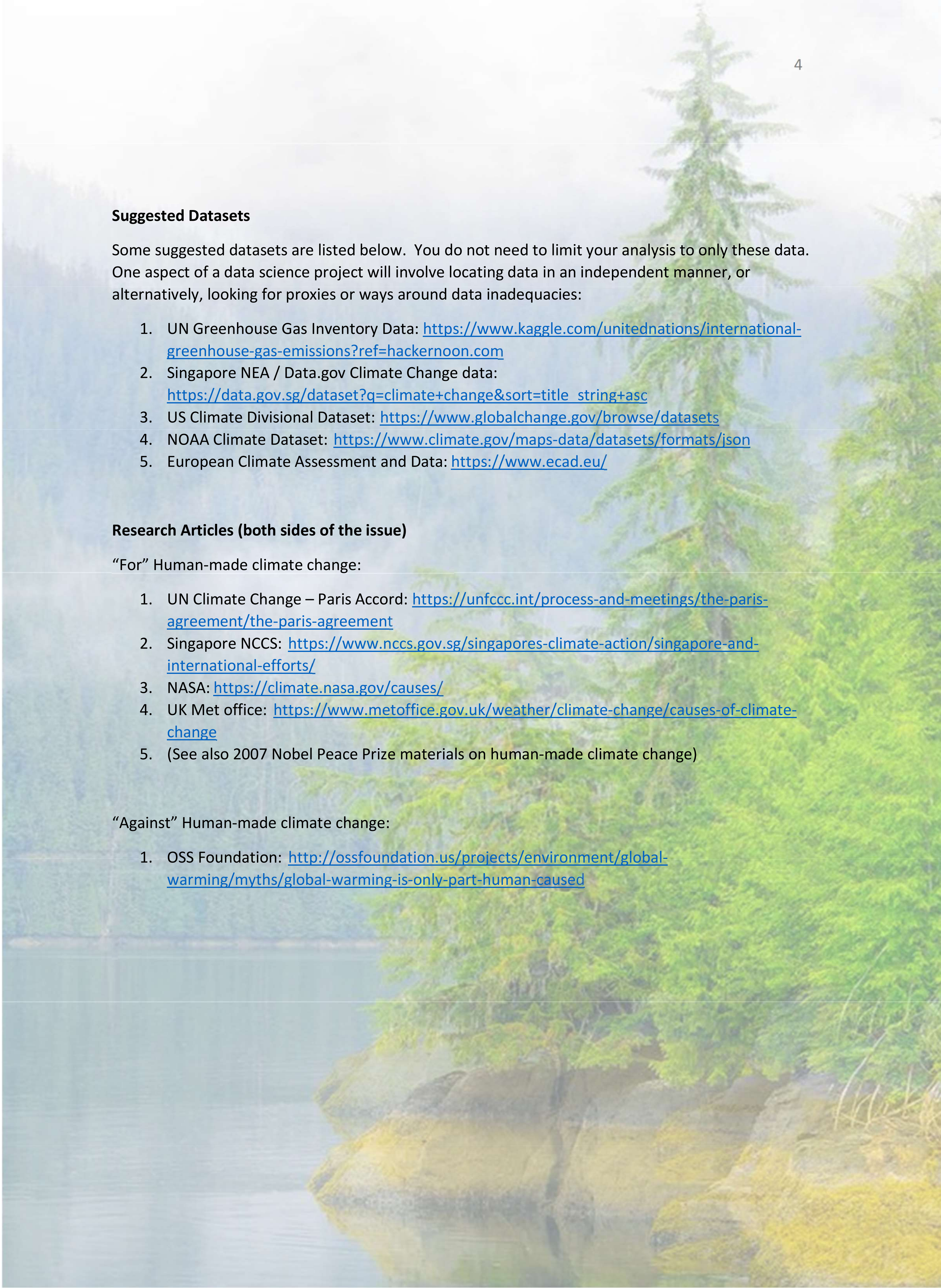
Source: Met Office
© FT



Sample project outline

1. Articulate scientific and economic intuition why there may be linkages between variables discussed above.
2. Are there univariate time series trends in the variables of interests?
3. Form and implement an explanatory model (e.g. VAR, ARIMA-X, ARMA-X) to determine if there is a robust econometric relationship between global temperature levels and greenhouse gases, as well as other potential explanatory variables. Also consider and test alternative explanations.
4. Explore relationships with secondary variables listed above, as well as (bonus) any additional variables of interest that you conceptualize separately

5. "Red Team Analysis" / Robustness Tests (Mandatory): Existing global opinion on "humanmade" climate change is almost, but not entirely unanimous. There are a subset of studies which claim that there is no statistically robust relationship between any form of human economic activity and climate change. Example is linked below. Replicate one or more of these studies which take the opposing point of view(s), and determine at a statistical level what are the causes of the differing conclusions. For instance, this could be different raw data, the variables used could have subtly different meanings, variation in statistical methodology (in this case: which approach is 'correct', are there any flaws, etc), differences in statistical significance criteria used, overfitting or different underlying economic or scientific assumptions.



Suggested Datasets

Some suggested datasets are listed below. You do not need to limit your analysis to only these data. One aspect of a data science project will involve locating data in an independent manner, or alternatively, looking for proxies or ways around data inadequacies:

1. UN Greenhouse Gas Inventory Data: https://www.kaggle.com/unitednations/international-greenhouse-gas-emissions?ref=hackernoon.com
2. Singapore NEA / Data.gov Climate Change data: https://data.gov.sg/dataset?q=climate+change&sort=title_string+asc
3. US Climate Divisional Dataset: https://www.globalchange.gov/browse/datasets
4. NOAA Climate Dataset: https://www.climate.gov/maps-data/datasets/formats/json
5. European Climate Assessment and Data: https://www.ecad.eu/

Research Articles (both sides of the issue)

"For" Human-made climate change:

1. UN Climate Change – Paris Accord: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement
2. Singapore NCCS: https://www.nccs.gov.sg/singapores-climate-action/singapore-and-international-efforts/
3. NASA: https://climate.nasa.gov/causes/
4. UK Met office: https://www.metoffice.gov.uk/weather/climate-change/causes-of-climate-change
5. (See also 2007 Nobel Peace Prize materials on human-made climate change)

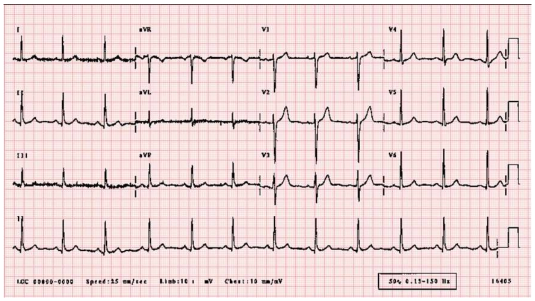
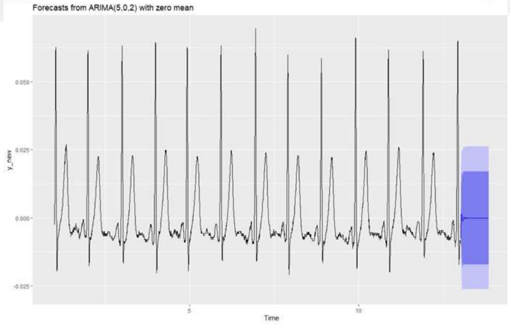
"Against" Human-made climate change:

1. OSS Foundation: http://ossfoundation.us/projects/environment/global-warming/myths/global-warming-is-only-part-human-caused

Option 2: Electro Cardiogram and Anomaly Detection

Electro Cardiogram (ECG) is used to measure strength of electrical currents flowing through heart muscles. We note that the human heart is composed of muscle tissues, which generally relies on regular, cyclical nervous signals (essentially weak electrical current) to contract and relax (to "beat" regularly). However, the heart is different from most other muscles in the human body because the nervous signals that cause it to beat comes not from the brain ("voluntary nervous system") but from the sinoatrial node, which is located in the heart itself. This is the natural "pacemaker" of the heart.

By measuring the strength of the electrical signals (using an "ECG") flowing through the heart, we are indirectly able to determine heart health. For instance, a sudden anomaly in the speed or amplitude of the ECG readings could indicate extremely rapid or irregular (or both) heart rate ("heart attack"). This might be seen as a change in the cyclical period of the ECG time series, or its absolute amplitude. On the other hand, a regularly occurring difference between an individual's ECG and that of the "baseline population" may indicate a chronic heart defect such as a valve anomaly, etc. One of the ways this could be modelled may be as statistically different ARMA coefficients in a statistical model of the heart.

Actual ECG (Source: ECGLibrary.com)	**ARIMA simulation of ECG (Source: RStudio):**
 I aVR V1 V4 II aVL V2 V5 III aVF V3 V6 II LOC 00000-0000 Speed:25 mm/sec Limb:10 mm/mV Chest:10 mm/mV 50~ 0.15-150 Hz 16405	

Sample project outline

1. Describe the underlying scientific relationships that link the ECG data to a diagnostic of heart health. Without looking at the data yet, form some simple hypotheses on different categories of heart health versus ECG readouts.
2. Visualize several ECG readouts, and discuss the best methodology for summarizing these in a time series model (e.g. ARMA, ARIMA, ARIMA-X, etc)
3. Using the suggested datasets, estimate "baseline" time series models for a "healthy human ECG", and as well as time series models for "distressed" ECG
4. Search for a model specification that produces the clearest separation in model estimates between "healthy / typical" and "distressed".
5. Robustness tests [Mandatory]: Consider "unusual" heart rates in otherwise healthy individuals. For instance, during cardio exercise. Is there a way to distinguish this from "unhealthy abnormal heart rates"? What about individual specific variation in what is healthy? For instance, infants or toddlers have higher resting heart rates. How would you incorporate these variation in the model?

6. Evaluate the algorithm as a medical diagnostic tool. What are the true positive / false positive and true negative / false negative rates?
7. [Remarks / Bonus]: Note that for step 4 above, machine learning is often used. This is formally beyond the coverage of this course. However, if you decide to try this based on expertise from other courses, ensure that in your presentation, you explain the machine learning framework in a simple intuitive way, (the explanation cannot rely on any technical terms or acronyms), so that the rest of your classmates can understand.

Suggested datasets:

1. PTB-XL (publicly available electrocardiography dataset) described on Nature: https://www.nature.com/articles/s41597-020-0495-6
2. Kaggle: https://www.kaggle.com/shayanfazeli/heartbeat

Research articles for background reading:

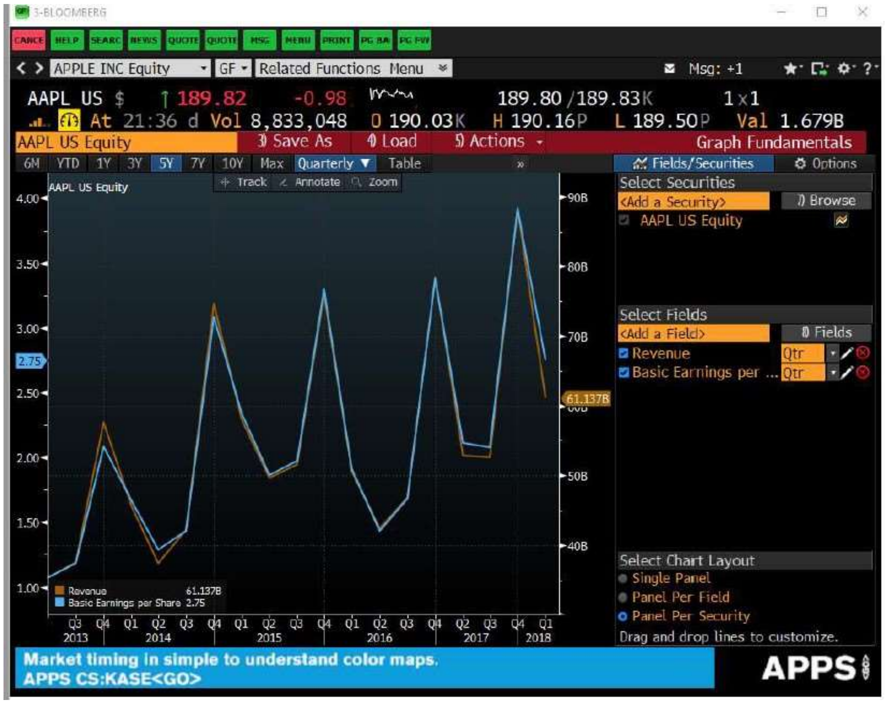
1. AR based method for ECG classification and patient recognition: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.736.4617&rep=rep1&type = pdf
2. Analysis and classification of heart diseases using heartbeat features [contains some machine learning content]: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0244-x

Option 3: Seasonality in Corporate Revenues

Corporate revenue performance often exhibits strong seasonality in some industries. For instance, in retail, there may be more expenditure towards the year end holiday gifting season. Major consumer goods corporations may also time their product cycle to release new versions towards the fourth calendar quarter (e.g. Apple has been releasing its new iPhones in September or October every year!). In recent years, corporations such as AliBaba or Amazon have also ‘invented’ and marketed their own versions of ‘public holidays’ such as ’11-11 Singles Day’ or ‘Cyber Monday’. All of these have a cumulative impact of increasing consumer discretionary expenditure (in aggregate) towards year end.

Apart from consumer spending, natural environmental factors (e.g. northern hemisphere winter) may also be a distinct contributory factor to seasonality. For instance, oil may exhibit greater demand in summer due to the “driving” and “travelling” season, while natural gas in winter due to heating purposes.

Effect might be confined only to some industry codes (“SIC codes”, or “Standard Industry Classification”). For instance, companies in the healthcare industry, utilities sector or consumer nondiscretionary may exhibit only limited seasonality relative to those more exposed to the effects discussed above.



AAPL US $ ↑ 189.82 -0.98 189.80 / 189.83K 1 x 1
At 21:36 d Vol 8,833,048 O 190.03K H 190.16P L 189.50P Val 1.679B

AAPL US Equity
3) Save As
4) Load
5) Actions
Graph Fundamentals

Select Securities
<Add a Security>
7) Browse
AAPL US Equity

Select Fields
<Add a Field>
8) Fields
☑ Revenue Qtr
☑ Basic Earnings per ... Qtr

Select Chart Layout
● Single Panel
● Panel Per Field
○ Panel Per Security
Drag and drop lines to customize.

Market timing in simple to understand color maps.
APPS CS:KASE<GO>

Overall, it might be interesting to forecast short term revenue increases accurately so that corporations can ensure sufficient finished product inventory to take advantage. Additionally, corporations may need to plan in advance for elevated staffing levels. Manufacturing firms with long supply chains could also need to alert their own raw materials or intermediate component suppliers. Lastly, this cyclical pattern in revenue figures may not be fully anticipated by public equity markets, resulting in a measure of short term return predictability.

Sample project outline:

1. Download quarterly revenue figures for all companies in the Russell 1000 from the SMU library's Compustat / CapitalIQ data subscription.
2. Using time series decomposition or other methods covered in class, determine which SIC codes typically have companies that exhibit strong seasonality, and which SIC codes only exhibit limited seasonality.
3. Articulate and discuss qualitatively (in as much detail as possible, preferably with multiple concrete examples at the company level, if not even more specific) the underlying economic reasons for the SIC codes that you have identified with 'strong' seasonality
4. Calibrate time series models (e.g. ARIMA-X) to quantify the seasonal revenue pattern of industries that you have identified with 'strong' seasonality. For simplicity, it may be easier to calibrate one model for each industry, rather than a single model across the entire equity universe. With reference to your discussion from step 3, elaborate on possible economic / structural reasons for any numerical differences in coefficient estimates.
5. [Robustness Tests]: In 2020 and 2021, it is possible that usual seasonal patterns may have been disrupted. For instance, air travel has been curtailed due to WHO pandemic recommendations. On the other hand, we may only have 6 or 7 quarters worth of data since any pandemic relevant restrictions were introduced; some are being, or have been since lifted. How would you determine (as best as possible) if the historical seasonal patterns that you identified previously still hold post pandemic?

Suggested datasets (all can be accessed via the SMU library):

1. Compustat
2. Capital IQ
3. Bloomberg

Research articles for background reading

1. Seasonality in the Cross Section of Stock Returns, Journal of Financial Economics: https://www.sciencedirect.com/science/article/abs/pii/S0304405X0700195X

Option 4: Forecasting Website Traffic

Being able to forecast incoming traffic to a website serves multiple purposes.

First, load balancing is a major factor in web services infrastructure. It ensures that website traffic are served by adequately provisioned servers. For instance, websites based on a cloud infrastructure can ensure that they ‘spin up’ additional compute nodes prior to any forecasted spike in inbound traffic

Second, website content updates or (conversely) maintenance cycles can be scheduled to take full advantage of peaks or troughs in anticipated traffic. Additionally, the impact of web-advertising can be better forecasted (and therefore, priced, in the case of fixed price advertising contracts) based on website traffic forecasts. Related to this, for web properties with multiple advertising contracts and with target obligations to meet quota for “number of impressions” by a certain date, the web property can better decide which advertisement to show at a certain time based on an accurate near term forecast.

Taking advantage of Wikipedia’s PageView API, which shows analytical data for article views on Wikipedia and its sister projects, we are able to construct time series models for website traffic based on various factors (e.g. time of day, weekend, holiday, as well as selected metadata on the articles and data on the incoming traffic such as country of origin, etc) that may be useful for forecasting

Option 5: Covid-19 infection forecasting

Infectious disease transmission is path dependent. The more individuals that catch an infection, the more new infections there will be in the future (ceteris paribus) because of a greater number of vectors

Against this backdrop, there are also relevant calendar effects (e.g. more infections on a weekend, public holiday, major event), and also the effect of treatment (such as social distancing, travel restrictions, no dining out, etc).

Combining these effects into a single time series model of new infection rate is interesting for two reasons. First, it allows us to better understand the “infectiousness” of the virus in the absence of any policy measures. Specifically, for each new infected individual, how many other individuals do they infect? The average effect of this can be estimated given time series on how infection numbers progress.

Second, the efficacy of policy measures can also be studied econometrically (similarly in terms of modifying the ‘infectiousness’ of the virus), and this may inform future policy updates in other situations (e.g. a new, previously undiscovered variant, or an unrelated new infectious illness where similar plans could be effective)