Industrial Problems Booklet

UNIVERSITY OF CAMBRIDGE
Faculty of Mathematics

Newton Gateway to Mathematics

Isaac Newton Institute for Mathematical Sciences

Smith institute

Contents

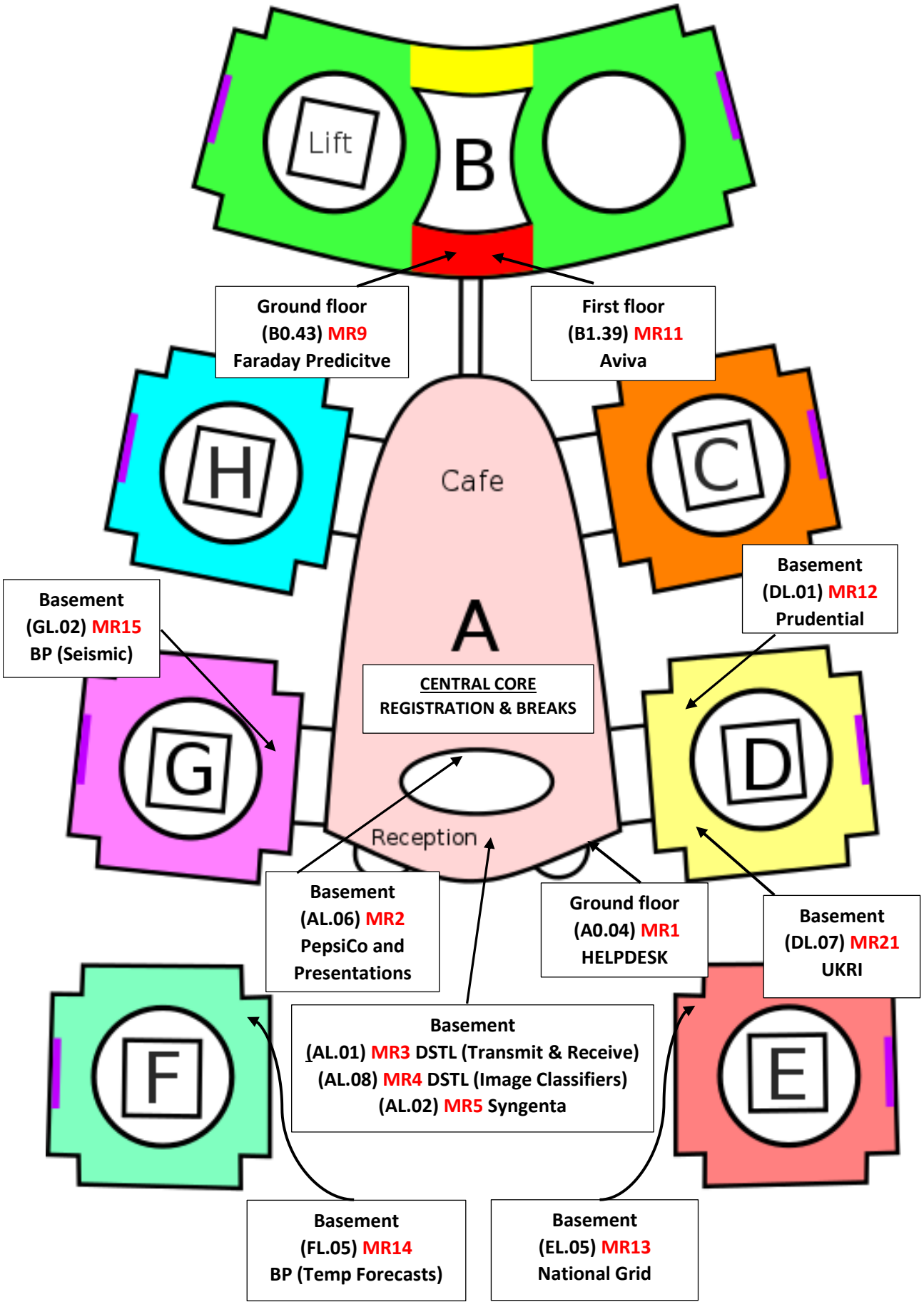
Page

Problems - Room Locations

Company	Problem	MR#	CMS#	Capacity
Aviva	Improving weather models for the insurance industry	MR11	B1.39	40
BP	Statistical modelling and pattern recognition for predicting evolution of temperature forecasts.	MR14	FL.05	50
Faraday Predictive	Identification of changes in noisy spectra – to detect incipient problems in rotating equipment.	MR9	B0.43	98
National Grid	The value of information in managing the electricity system	MR13	EL.05	50
Prudential	Quantile regression with a time series structure	MR12	DL.01	40
DSTL	Limits on simultaneous transmit and receive	MR3	AL.01	108
BP	Uncertainty in seismic inverse problems.	MR15	GL.02	46
PepsiCo	Mash disc forming	MR2	AL.06	180
Syngenta	Towards managing landscapes: how can we interpret and design better environmental monitoring surveys?	MR5	AL.02	60
DSTL	Hardening techniques for image classifiers.	MR4	AL.08	60
UKRI	Industrial Strategy Challenge Fund Discussion (Wednesday 15:45-17:30)	MR21	DL.07	25

Confidential Disclosure Agreements (CDAs)

The following companies have CDAs: BP (both problems), National Grid and Prudential. These cover confidential code, algorithms, information or data that the Company may share with the group. The University of Cambridge has signed the CDAs, which will cover the ESGI delegates, so there is no need for individual delegates to sign. However, please note that delegates who download or use the confidential code, algorithms, information or data will by that action be agreeing to be bound by the CDA. The CDAs are on display in the relevant rooms.


Lift
B
Ground floor
(B0.43) MR9
Faraday Predicitve
First floor
(B1.39) MR11
Aviva
H
Cafe
C
A
Basement
(GL.02) MR15
BP (Seismic)
Basement
(DL.01) MR12
Prudential
CENTRAL CORE
REGISTRATION & BREAKS
G
D
Reception
Basement
(AL.06) MR2
PepsiCo and
Presentations
Ground floor
(A0.04) MR1
HELPDESK
Basement
(DL.07) MR21
UKRI
F
Basement
(AL.01) MR3 DSTL (Transmit & Receive)
(AL.08) MR4 DSTL (Image Classifiers)
(AL.02) MR5 Syngenta
E
Basement
(FL.05) MR14
BP (Temp Forecasts)
Basement
(EL.05) MR13
National Grid

Monday – Problem Introductions

Problem Sessions

09:40-10:05	Aviva Improving weather models for the insurance industry Mick Comerford
10:05-10:30	BP Statistical modelling and pattern recognition for predicting evolution of temperature forecasts Milos Krkic
10:30-10:55	Faraday Predictive Identification of changes in noisy spectra – to detect incipient problems in rotating equipment Geoff Walker
10:55-11:20	Morning Coffee Central Core, CMS
11:20-11:45	National Grid The value of information in managing the electricity system Andrew Richards
11:45-12:10	Prudential Quantile regression with a time series structure Ziwei Wang and Daniel Slavik
12:10-12:35	DSTL Limits on simultaneous transmit and receive Christopher Swinerd
12:35-13:00	BP Uncertainty in seismic inverse problems York Zheng
13:00-14:00	Lunch Central Core, CMS
14:00-14:25	PepsiCo Analysis of shear forces during mash disk formation Tom Bullock
14:25-14:50	Syngenta Towards managing landscapes: how can we interpret and design better environmental monitoring surveys? Paul Sweeney
14:50-15:15	DSTL Hardening techniques for image classifiers Phillippa Spencer and Sophie Debenham

Friday – Final Presentations

Presentation Sessions

09:00-09:25	Aviva Improving weather models for the insurance industry
09:25-09:50	BP Statistical modelling and pattern recognition for predicting evolution of temperature forecasts
09:50-10:15	Faraday Predictive Identification of changes in noisy spectra – to detect incipient problems in rotating equipment
10:15-10:40	National Grid The value of information in managing the electricity system
10:40-11:00	Morning Coffee Central Core, CMS
11:00-11:25	Prudential Quantile regression with a time series structure
11:25-11:50	DSTL Limits on simultaneous transmit and receive
11:50-12:15	BP Uncertainty in seismic inverse problems
12:15-12:40	PepsiCo Analysis of shear forces during mash disk formation
12:40-13:20	Lunch Central Core, CMS
13:20-13:45	Syngenta Towards managing landscapes: how can we interpret and design better environmental monitoring surveys?
13:45-14:10	DSTL Hardening techniques for image classifiers
14:10-14:20	Conclusions and Awards Depart

Improving weather models for the insurance industry

Introduction

Predicting the risk of extreme weather events, such as floods, storms, or ground frost is critical in property insurance. The dominant approach in the primary insurance industry is to classify the risk of events into coarse categories — a 1-in-100 year or 1-in-250 year storm or flood, for example — without time dependence. This ignores the effect of global trends which could influence the future activity of extreme events on a range of time scales.

The broad goal for the study group is to evaluate the feasibility of developing more detailed estimates of extreme weather risk. These could allow an insurance company to:

- Invest in claims, underwriting and pricing initiatives to diversify risk
- Guide reinsurance strategy
- Inform customers on how best to prepare during a high-risk period
- Ensure that claims departments have all necessary resources in place
- Plan early interventions to minimise damage in the event of a storm, flood, or freeze.

Specific goals

1) To propose strategies for estimating the risk of extreme weather in the UK seasons to decades ahead
2) To explore the seasonal dependence of weather risks, using historical data.
3) To evaluate the effect that global phenomena such as climate oscillations, global warming, and even solar activity may have on risk estimates.
4) To suggest metrics for the assessment of the benefits of new weather risk estimates using insurance claims data.

Methodological considerations

There are various ways of estimating risks of weather events that can cause severe damages and losses. Historical data can be analysed to determine frequency of occurrence and probability distributions — but data may be sparse and may have trends. *Weather generators* (statistical models fitted to past time series) can be used to simulate weather and create a wide range of scenarios. There are several methodological challenges to consider: (i) how best to deal with correlations between weather variables such as precipitation, temperature, and frost, (ii) how to model the spatial

dependence of the variables, (iii) how to model the probability distribution at each time point and the risk of extreme events.
Dynamical models like those used for numerical weather forecast systems and climate change projections provide outlooks for days to decades ahead, and data from such systems may be used to modify risk estimates. The climate system contains slow variability on a wide range of timescales, and knowledge of the observed current climate state can condition risk estimates.

Participants should consider a range of approaches to estimate relevant risks. It will be important to suggest methods to cross-validate the model with the data available and evaluate the predictive value of different modelling choices.

More broadly, participants can consult experts from Aviva to further define the scope of the problem. For example, what is the geographic area of interest; how far in advance are estimates useful, and what types of extreme events may be of interest? An important part of the exercise will be to explore what data are needed to train a model, what is available, and consider how to acquire more data relevant to the problem – observations, simulations, and forecasts.

Finally, it is important to consider how to validate the weather risk estimates against insurance claims data, combining the predictions with important side information available to the insurer, such as detailed maps of flood risk.

Data and modelling resources

There are many relevant sources of data, model, and risk information, and an initial task will be to explore these. A few starting points are provided here.

Observational data: The Met Office provides open access to several observational datasets. In particular, the HadUK Gridded Dataset contains data from weather stations across the UK going back to 1862 interpolated on grids of various sizes over the country. This dataset contains information on air temperature (monthly means, minima, and maxima), precipitation, wind speed, and days of ground frost, among others.

A broader source of climate datasets including observational data, seasonal forecasts, and climate change projections is provided by the Copernicus Programme.

The IRI Climate Data Library is an alternative source for climate datasets, projections, and analysis tools.

Long-range forecasts:
https://climate.copernicus.eu/seasonal-forecasts
https://www.metoffice.gov.uk/research/climate/seasonal-to-decadal/long-range/index

Stochastic weather generators:
https://www.ipcc-data.org/guidelines/pages/weather_generators.html
https://rmets.onlinelibrary.wiley.com/doi/10.1002/joc.3896

Return periods:
http://climatica.org.uk/climate-science-information/return-periods-extreme-events
https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.1155

Climate oscillations: Climate oscillations are recurrent patterns with wide-ranging, effects on e.g. temperature, precipitation, and storm tracks. They are irregular, although some oscillations have a

characteristic time scale. In particular, the North Atlantic oscillation has an important effect on UK weather.

North Atlantic Oscillation
https://en.wikipedia.org/wiki/North_Atlantic_oscillation
https://www.cpc.ncep.noaa.gov/data/teledoc/nao.shtml
https://www.metoffice.gov.uk/research/climate/seasonal-to-decadal/gpc-outlooks/ens-mean/nao-description

Atlantic Multidecadal Oscillation
https://en.wikipedia.org/wiki/Atlantic_multidecadal_oscillation
https://climatedataguide.ucar.edu/climate-data/atlantic-multi-decadal-oscillation-amo

Madden-Julian Oscillation. 'The 40-day wave' mainly impacts the tropics, but also has mid-latitude effects.
https://www.metoffice.gov.uk/weather/learn-about/weather/atmosphere/madden-julian-oscillation
https://www.climate.gov/news-features/blogs/enso/what-mjo-and-why-do-we-care

Climate change and extremes:
https://www.ipcc-data.org/sim/ar5_tables/ar5_extremes.html

Solar activity: A relationship between solar activity and the climate has been previously reported. For an example, see
https://www.nature.com/articles/ngeo1282

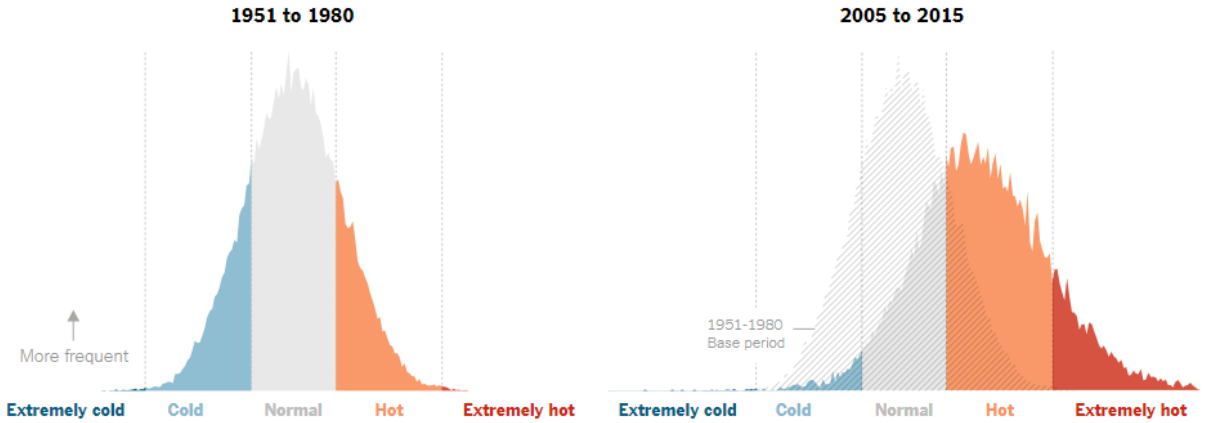
Statistical modelling and pattern recognition for predicting evolution of temperature forecasts

Company Background

BP is a global energy company that is leading the transition to a lower carbon future. No matter your role here, we each play a part in progressing the way we deliver heat, light and mobility to the world. Our business is the exploration, production, refining, trading and distribution of energy. BP operates with business activities and customers in more than 80 countries across six continents and every day we serve millions of customers around the world. Integrated Supply & Trading (IST) is BP's commercial face to the global energy and commodity markets. We market the company's upstream production of hydrocarbons, secure feedstocks for our refinery system and provide services to external customers, including fuel supply and hedging solutions.

Problem Background

Global energy companies deeply care about the weather in order to forecast energy demand for most of Europe and North America. In the US, the energy demand peaks twice a year driven by heating demand in winter months and power demand in summer months. With extreme weather becoming the norm especially during summer months (see the chart below that depicts how summer months have become more extreme globally), energy companies such as BP need to pay a close attention to not only the current forecast of temperatures but also to how the forecast may evolve in time (known as the forecast of forecast).



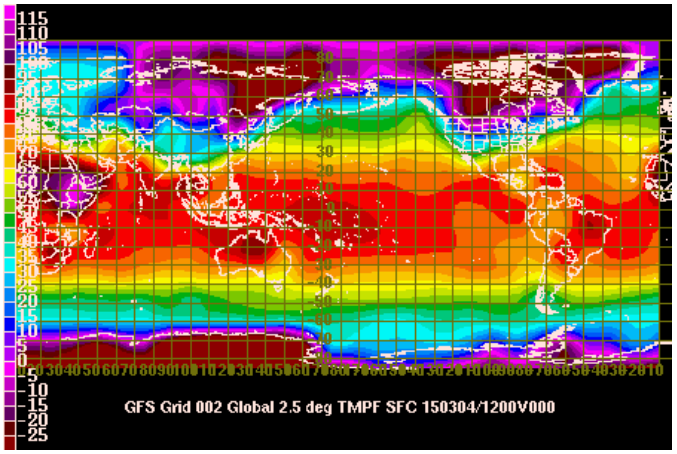
Source: NY Times

Currently there are two global numerical weather prediction (NWP) systems relied upon for temperature forecasts by academics, industry, enthusiasts and the governments across the world. These are –

1. Global Forecast System (GFS) – Run by the US National Weather Service, this model runs 4 times a day and produces forecasts for up to 16 days in advance at varying spatial resolution (13 km to 27 km). This model is a basis for derivative models such as Global Ensemble Forecast System (GEFS) and North American Ensemble Forecast System (NAEFS).
2. ECMWF European Model – Run by European Centre for Medium-Range Weather Forecasts (ECMWF), this model also runs 4 times a day. Like GFS, there are derivative models run by the agency with varying resolution to balance computational cost and accuracy.

Data

The numerical models (GFS or ECMWF) divide the globe into a grid and predict temperatures for the grid points at regular time increments out the next 15 days. Shown below is a snapshot of the output from GFS model.



For our use cases, temperature data is condensed into Average Daily Temperature representing a single value for each location and for each day.

For the scope of this project, you will be provided with 15-day temperature forecast for the period of 1-Mar-2007 to 28-Feb-2019 for 150 locations that fall within the continental US. This dataset is sourced from a vendor that uses proprietary algorithms to average and bias-correct the intraday outputs of GFS and ECMWF weather runs along with additional human forecaster nudging at times, resulting in a single forecast vector for each day. A sample from the dataset is shown below.

RUN_DATETIME	LOCATION_CODE	OBSERVATION_DATE	TEMPERATURE
3/1/2007	1	3/1/2007	54.5
3/1/2007	1	3/2/2007	47
3/1/2007	1	3/3/2007	40.5
3/1/2007	1	3/4/2007	37
3/1/2007	1	3/5/2007	40.5
3/1/2007	1	3/6/2007	44.5
3/1/2007	1	3/7/2007	48.5
3/1/2007	1	3/8/2007	51.5
3/1/2007	1	3/9/2007	54.5
3/1/2007	1	3/10/2007	52
3/1/2007	1	3/11/2007	49.5
3/1/2007	1	3/12/2007	51.5
3/1/2007	1	3/13/2007	54.5
3/1/2007	1	3/14/2007	54.5
3/1/2007	1	3/15/2007	52.5
3/2/2007	1	3/2/2007	47.5
3/2/2007	1	3/3/2007	42.5
3/2/2007	1	3/4/2007	37
3/2/2007	1	3/5/2007	41.5
3/2/2007	1	3/6/2007	46
3/2/2007	1	3/7/2007	50.5
3/2/2007	1	3/8/2007	53
3/2/2007	1	3/9/2007	53.5
3/2/2007	1	3/10/2007	51.5
3/2/2007	1	3/11/2007	52.5
3/2/2007	1	3/12/2007	53.5
3/2/2007	1	3/13/2007	56.5
3/2/2007	1	3/14/2007	54.5
3/2/2007	1	3/15/2007	52.5
3/2/2007	1	3/16/2007	50.5

1. RUN_DATETIME: The ‘as of date’ for the forecast
2. LOCATION_CD: Location code
3. OBSERVATION_DATE: The forecast target date
4. TEMPERATURE: Average Daily Temperature for the location

Problem Description

Each day's forecast at a given point is comprised of 15 numbers which represent temperatures for days 0-14 in the future. As we move to the next day the forecast vector rolls by 1 day such that tenor 14 of previous day's forecast now becomes tenor 13 of current day's forecast. Predicting the weather forecast for the next day is to predict the 15-dimensional vector for the next day. The table shown below illustrates the structure of the data.

	TEMPERATURE																				
OBSERVATION_DATE	3/1/2007	3/2/2007	3/3/2007	3/4/2007	3/5/2007	3/6/2007	3/7/2007	3/8/2007	3/9/2007	3/10/2007	3/11/2007	3/12/2007	3/13/2007	3/14/2007	3/15/2007	3/16/2007	3/17/2007	3/18/2007	3/19/2007	3/20/2007
RUN_DATETIME																				
3/1/2007	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x					
3/2/2007		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x				
3/3/2007			x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			
3/4/2007				x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
3/5/2007					x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
3/6/2007						x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Problem Statement

The problem we present to you is as follows: estimate, for each time point t, the conditional distribution of the 150 × 15-dimensional vector corresponding to the 15-day forecast at each location, given the historical data up to time t-1.

Note: datasets are provided to the working group under a Confidentiality Agreement signed on behalf of the University of Cambridge.

Identification of changes in noisy spectra – to detect incipient problems in rotating equipment

Background/overview of problem context

Faraday Predictive is a small Cambridge-based technology company, specializing in the predictive maintenance of rotating industrial equipment such as pumps, fans, compressors, and conveyors. Issues associated with maintenance of this industrial equipment are estimated to cost $700 billion pa worldwide, with much of this money being wasted through inappropriate maintenance strategies. Faraday Predictive provides a means of remotely monitoring rotating equipment and diagnosing impending faults. This helps the customer (who might be a water company, for example) maintain their assets in a timely manner and avoid a catastrophic machine failure by scheduling preventative actions well in advance and conversely to avoid doing un-necessary maintenance on a time-schedule when it is not required.

Our technology uses the electric motor driving the equipment as a sensor, by measuring the voltage applied to, and current drawn by, the motor, and identifying subtle distortions in the shape of the current waveform relative to the voltage waveform. These relative distortions, identified by means of a mathematical modelling approach, are expressed as a residual current. The frequency components in this residual current correspond to the characteristic frequencies of the phenomena causing them, which are typically related to deterioration of the equipment, such as bearing wear, belt slippage, internal corrosion, rubbing, misalignment, etc. By matching the observed frequencies against known characteristic frequencies, we are able to identify the likely cause of the distortion, and the amplitude of the signal at this characteristic frequency indicates the severity of the problem.

The overall steps in this process are as shown in figure 1 below:

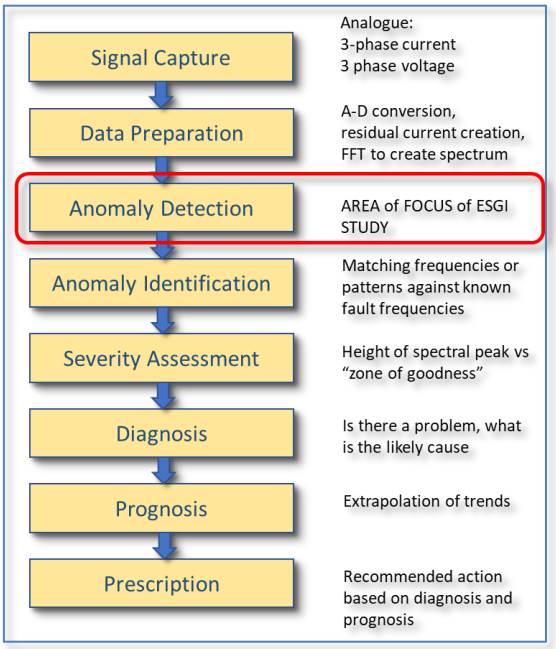


Figure 1 - overview of mathematical processes

Problem Briefing

A range of mathematical techniques are used in each step of this process, and we can describe them and explore limitations or alternatives to them if this is of interest, particularly if the initial "exam question" is solved early in the week. However, the "Exam Question" is a specific one in the area of Anomaly Detection:

- **How can we reliably detect changes of shape of our spectrum given a noisy signal?**

Once this step is firmly established, the subsequent steps in the process can be called into action, and whilst we can describe and discuss these other steps with the group, their effective deployment is all predicated on having detected the anomaly in the first place, so that is where we want the group to focus first.

For some failure modes, the problem manifests itself as a simple peak at a particular location on the spectrum, and this can be detected relatively easily. This is not the focus for ESGI.

However, some other phenomena show up as a broad "hump" of signals rather than at a single peak, and so far we do not have a good solution to identifying this sort of shape change. Examples of this issue are shown in figures 2 & 3 below:

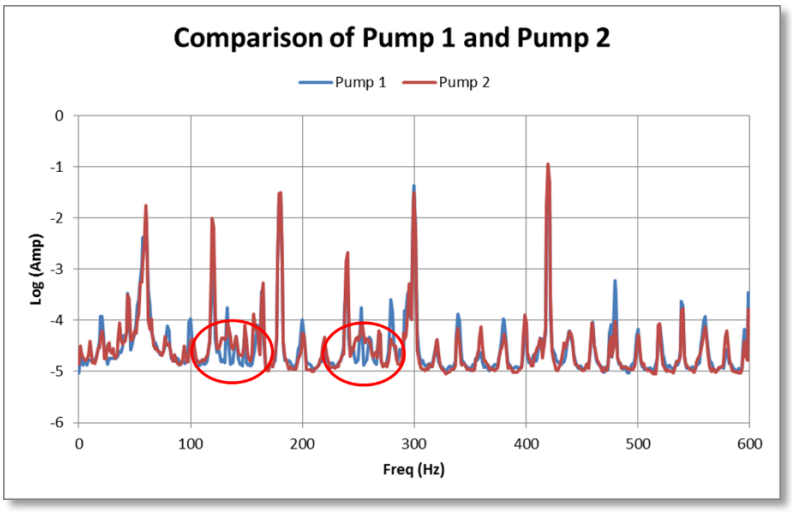


Figure 2 - shows the spectra for two identical pumps - with the red one showing two elevated areas

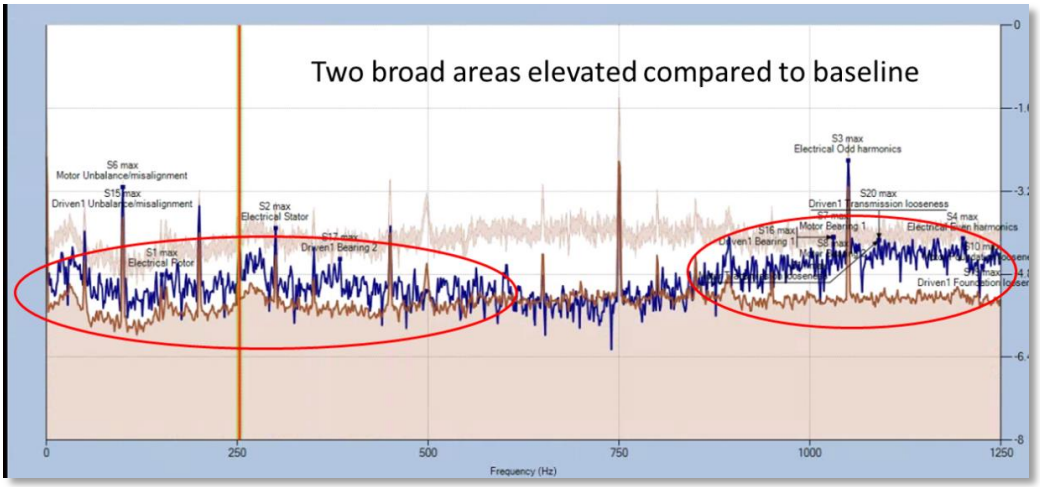


Figure 3 - Spectra for one item of equipment that has changed over time. Brown = baseline from October, Blue = single instant in December (note shape continues to display this elevated shape over extended period, eg right up to Feb 2019)

Figure 2 shows the spectra for two otherwise identical pumps. It can be seen how similar the shapes are, which indicates how similar the pumps are and what similar duties they are on. However, it is clearly visible to the eyeball that there are two distinct humps on the red trace that are not present on the blue trace. It is believed that these humps indicate the early stages of deposit build up inside the pump. What we seek is an automated method for spotting this sort of anomaly.

Figure 3 shows two spectral traces – the blue one is a single "instant" spectrum corresponding to a measurement made at a particular point in time. The brown trace is a baseline, created by combining and averaging a number of consecutive spectra at an earlier period. Two points to note from this figure are firstly, the instant spectrum is much noisier than the baseline, where noise has been averaged out; and secondly, the two broad areas where the more recent blue spectrum is elevated compared to the baseline, indicating something is going on, ie some deterioration is occurring in the rotating equipment. We seek a way of automatically alerting users to the presence of this change, without creating false alarms from random noise.

If the primary “exam question” were solved early on during the week, a number of other areas would also be of interest, including:

- What causes these humps, as opposed to peaks?
- What causes signals to show up as sidebands on higher harmonics as opposed to sidebands on the fundamental? In addition, is there a rational basis for weighting the significance of one of these higher harmonic peaks as compared to ones on the fundamental?
- Empirically, we see subharmonics (e.g. 1/3, 1/5, 1/7th and sometimes multiples of these, e.g. 2/7th, 3/7ths) of the rotational speed when rubbing friction is present. Can you explain why this should be the case, and why it shows up at the particular frequency in any particular case (e.g. why sometimes 1/5, and other times, 1/7th)?

Available Data and Tools for the study group

We can make available a number of data sets that are held in SQL databases, and tools that allow easy viewing of the spectra, trends, and in some cases the underlying source voltage and current waveforms.

nationalgrid

The value of information in managing the electricity system

Introduction

The GB electricity network is organised into the Transmission network and Distribution networks. Distribution networks are in charge of all the distribution within their local area (a local area could be London or South-West England for example). The Transmission network is a high voltage network designed to transport electricity over large distances. Distribution networks connect to the Transmission network at suitable points. As a schematic representation, the Transmission network can be viewed as a central hub, with distribution networks attached radially (see Fig. 1).

Electrical power is provided by generators. Large generating units (shown in Fig. 1 by the green boxes) are known as Balancing Mechanism Units (BMUs) and these units connect either directly to the transmission network or to their local distribution networks. Small generating units, termed non-BMUs, connect directly to their local distribution networks. These latter generators may be solar panels on people's houses, wind farms or local diesel generators/batteries (shown in the Fig. 1 by the red circles). This means that at any given time some of GB's total electricity demand is met by BM units and some by non-BM units.

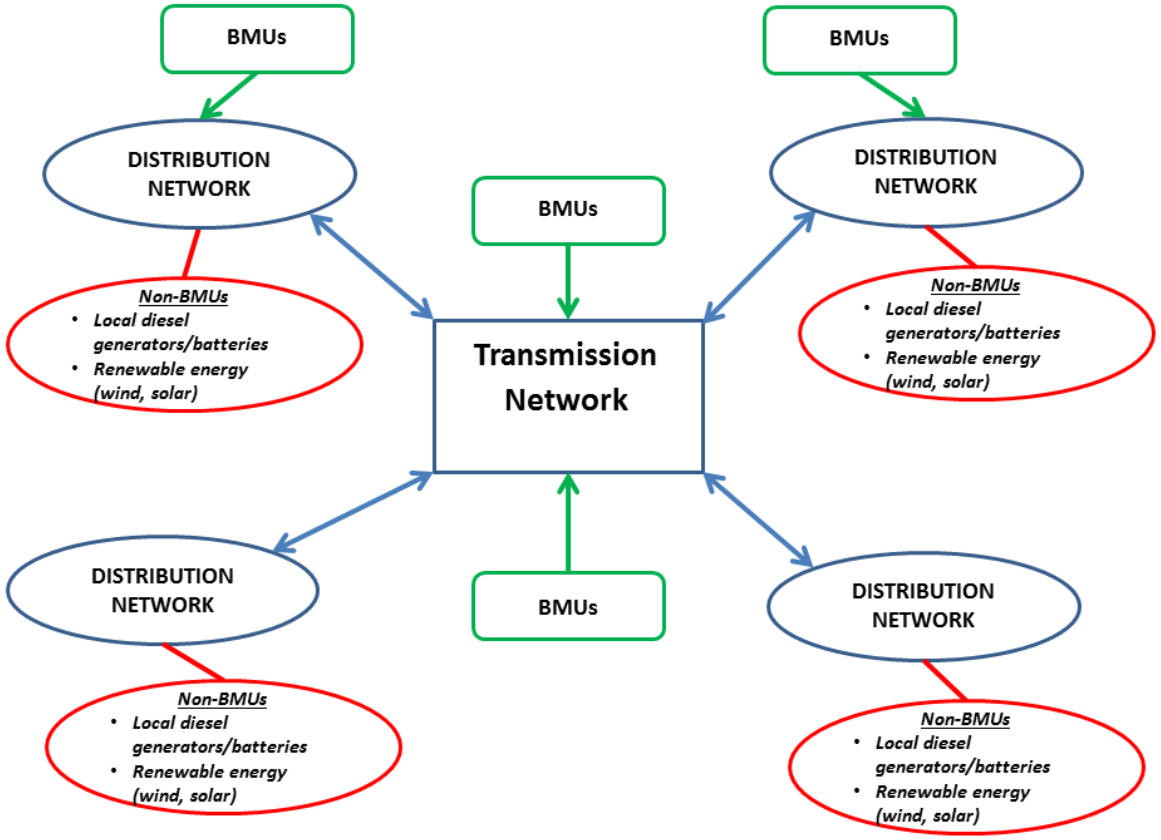


Figure 1: Approximation of the electrical network in Great Britain.

The Transmission System Operator, TSO (in Great Britain this is National Grid Electricity System Operator, NG ESO) has to manage the entire system. It is its job to keep the total supply of electrical power and the total demand throughout the country perfectly in balance at all times. The TSO does this by forecasting the supply-demand balance so that it can accurately schedule generation (this has to be done ahead of real time). The TSO also needs to schedule a certain amount of 'reserve' generation which can be used at short notice in order to meet an unexpected increase in demand or sudden loss of scheduled generation. This adds considerably to the expense of running the system.

However, the problem now faced by the TSO is that it is only aware of – and can therefore only control – the generation being provided by the BMUs. These provide generation schedules ahead of time and real-time output data to the TSO. Therefore, the BMUs are visible to and controllable by the TSO. On the other hand non-BMUs (all connected directly to the distribution networks) do not provide such information to the TSO, and indeed the TSO in general does not know how many non-BMUs there are, their capacity, location and fuel type. Therefore, the non-BMUs are invisible to the TSO and the only way they are 'seen' is as a net change in the total demand on the system. The activities of the non-BMU generators, and in particular the fact that these activities are not known to the TSO, therefore add greatly to the uncertainties involved in managing the system.

Traditionally the impact of non-BMUs on the system has been small. However, the rise of renewable energy, such as wind and solar, has led to a huge increase in non-BMUs. As a result there is now much greater variability in the TSO's forecast of the supply-demand balance (since non-BMUs are invisible to the TSO) and these forecasts are becoming less and less accurate. This in turn means that greater levels of reserve generation must be scheduled, thereby considerably increasing the total cost of running the system.

If there were more information on the activities of non-BMUs this could be incorporated into the TSO's forecasting models and the TSO would then be able to make more accurate forecasts and reduce the required levels of reserve, thus reducing costs. But the question remains - what is the value of this additional information on non-BMUs and how much could this information improve current forecasts?

Problem Question

What is the value of greater knowledge about the activity of non-BMUs versus the cost of needing increased reserves to cope with the increased variability in forecast supply-demand balances?

Data that you might find useful to answer this question are available at -
https://www.nationalgrideso.com/balancing-data/data-explorer

Note: datasets are provided to the working group under a Confidentiality Agreement signed on behalf of the University of Cambridge.

PRUDENTIAL

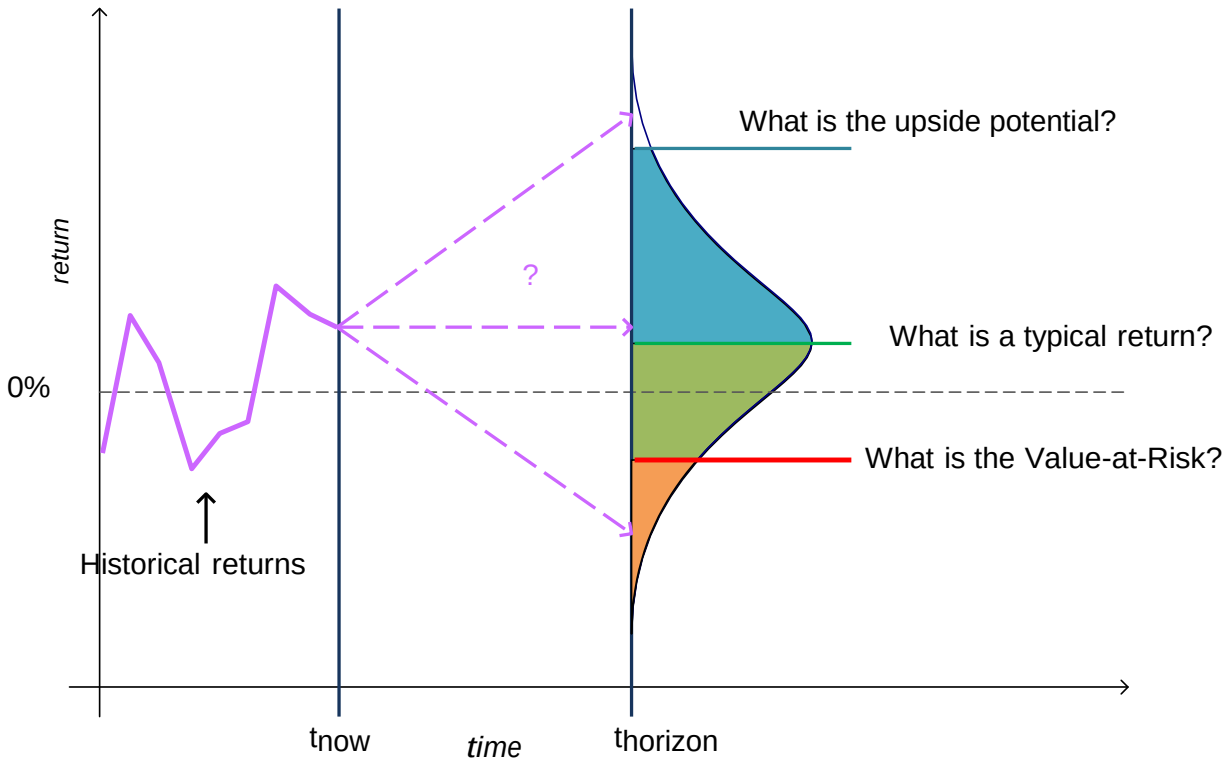
Conditional quantile estimation using high-dimensional time series data

Motivation

Stock market returns are subject to large drawdowns, which can take years to recover. The global financial crisis led to a fall of $7.9 trillion in the market capitalisation of the S&P500. It took investors over 3 years to recover their capital, assuming they stayed fully invested: many do not. The fallout from the dot-com bubble took even longer to recover. Avoiding large drawdowns is crucial for savers and pensioners. It can have a dramatic effect on the ending wealth of their portfolios. Abstract movements in stock markets are life-changing events.

As asset allocators, the view of future returns is a crucial input in securing solid performance and avoiding large drawdowns. The challenge is not just to understand the median or mean return but the distribution of returns and the size and probability of tail risk. The problem we outline below is a significant part of this challenge.

Figure 1 Conditional quantile estimation of asset returns



The problem

An important problem faced in our long-term strategy work is to deal with estimating the broad future outcomes of time series, with a slightly greater emphasis on extremes. In this instance, the preference is to focus on estimating conditional quantiles of the series, given other historical data, rather than estimating a generative parametric model for the data.

We have available data consisting of monthly and quarterly observations of around 50 continuous economic and financial variables covering the period since 1970. We also have available a further time series over the same period to be treated as a response. Our goal is to estimate, for each time point, the conditional quantiles of the response series given historical observations of the remaining series.

Our problem may differ from classical forecasting problems in the following aspects:

- We are interested in estimating conditional quantiles rather than conditional means.
- We have a large number (around 50) of covariate series we would like to condition on instead of conditioning on the history of the response series itself. Moreover, we have relatively few observations (around 600 for the monthly series) making our problem high-dimensional.
- We are interested in exploring methods that do not make strong parametric assumptions on the time series.

Questions

Q1: Given the nature of the problem described above, which methods are particularly suitable to attack this problem?

Q2: What would be a good validation framework allowing us to compare performance of various conditional quantile estimation methods (existing versus new methods)?

Q3: How robust / stable are the proposed methods in terms of their sensitivity to assumptions, and can one provide guarantees for the estimated conditional quantiles in terms of confidence intervals?

Q4: So far, the problem has focused on the conditional distribution of a single response variable. How could one extend the methods to cater for joint conditional distribution estimation when we have a vector of time series?

The available dataset

Whilst some of the economic time series were designed with stationarity in mind, other time series will have to be transformed to be approximately stationary.

An important aspect to consider is avoiding look-ahead bias: when estimating the conditional quantile of interest, only sufficiently lagged observations can be used (for instance, end-of-quarter data are typically released few months later). Look-ahead bias must be also carefully reflected in cross-validation (CV) set construction/generalisation error estimation.

Note: datasets are provided to the working group under a Confidentiality Agreement signed on behalf of the University of Cambridge.

Limits on Simultaneous Transmit and Receive

The Challenge

To understand the fundamental limits on our ability to emit and receive radio signals at the same frequency and at the same time (i.e. in same-channel full duplex mode) using a transceiver radio system.

For this challenge we want to focus not on improving the technology but on finding the fundamental mathematical limits on what can be achieved — limits that will apply whatever technology is used: limits perhaps analogous to Cramer-Rao lower bounds on variance, or the Shannon-Hartley limit on information rate.

Overview

The ability to Simultaneously Transmit and Receive (STAR) signals in the radio band of the Electromagnetic (EM) Spectrum (EMS) offers huge benefits to civil and defence applications. Current approaches[1] achieve this through time ***or*** frequency division duplexing. However, same-channel full-duplex (i.e. working simultaneously in frequency ***and*** time) offers huge efficiencies and system performance benefits compared to frequency or time–division duplexing.

The current state-of-the-art in academic and industrial STAR research shows that ~120dB of isolation depth can be achieved in a relatively narrow operating bandwidth. Whilst we seek to extend both isolation depth and operating bandwidth (our current aspirations are for 150dB of isolation depth over 160MHz of operating bandwidth), we need to understand what the theoretical limits are in order to direct our research investment.

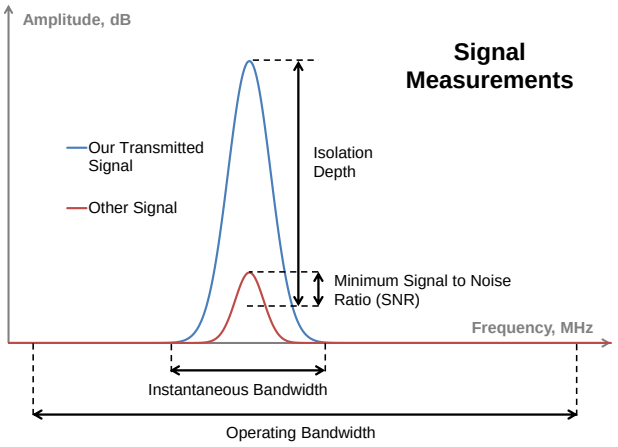


Figure 1: - Diagram of Signal Measurements

There is some analogy here to noise cancellation, such as used in commercial noise-cancelling headphones. However, in this case we generate the primary interferer (the noise) that we want to cancel. Cancellation performance will, therefore, be limited by our ability to accurately model our own emissions and the changes that happen to them after they are transmitted into the Electromagnetic Environment (EME).

[1] A Survey of Self-Interference Management Techniques for Single Frequency Full Duplex Systems - Nwankwo 2017 - http://eprints.gla.ac.uk/151582/7/151582.pdf

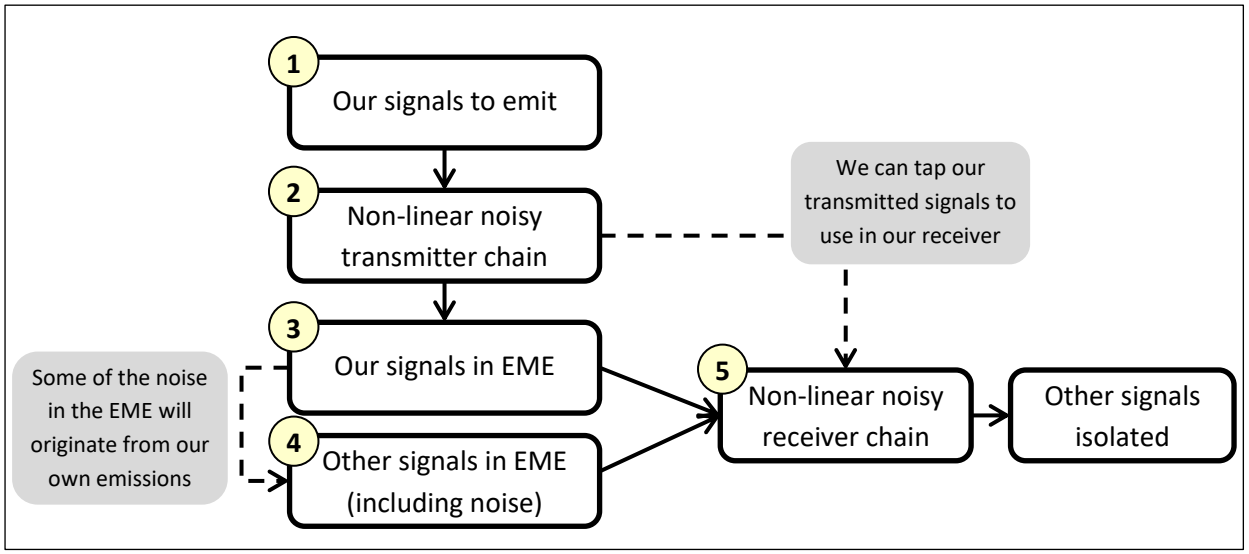


Figure 2: Flow model for the challenge

A simple information flow model for this challenge is presented at Figure 2. It shows:

1. **Our Signals** - that we wish to emit to the EME.
2. **Non-Linear Noisy Transmitter Chain** - our signals pass through this to enter the EME.
3. **Our Signals in EME** - Once in the EME, our emitted signals can couple or propagate to our receiver chain (coupling and propagation is used here to differentiate between energy coupling due to near-field or far-field EM physics, respectively) along with any other signals, including noise, that are present in the EME.
4. **Other Signals in EME** – As well as our own emissions there may be other signals in the EME (the signals we want to detect)
5. **Non-Linear Noisy Receiver Chain** - All the signals pass through our own noisy receiver chain from which we wish to detect the presence of other signals: signals that are at the same frequency as our own emitted signals at the same time. We are able to tap signals off our noisy transmitter chain at any stage and use this information within our receiver chain.

The references to noise in Figure 2 highlight that the knowledge of our transmitter and receiver chains and of the EME may not be perfect. For example, although we know the properties of the signal that we emit it will pass through the EME before it reaches our receiver. In the environment it will experience non-linear effects (i.e. reflections, diffraction, attenuation, signal conversion) that mean what is received is not identical to what we emitted. We need to be able to model these effects to be able to cancel out our signals and detect the other signals that we are interested in. The limits of the model will infer the limits of our capability.

Physical Aim	Design Approach
1. EM Protection layer – Stop the emitters 'frying' the sensors.	E.g. Antennas, circulators - high transmit to receive isolation
2. EM Linearity layer – Maintain the sensor systems in linear region.	Analog processing e.g. Interference cancellation, (passive or active).
3. EM Own signals layer – See through own platform emissions.	Digital processing e.g. signal cancellation, adaptive isolation
4. EM Known signals layer – Seeing through known signals in environment.	Digital processing e.g. coded cancellation, side-channel cancellation

Figure 3: Layered model for STAR

An engineering approach to exploiting the information flow within this challenge, as outlined at Figure 2, is expressed at Figure 3. This model highlights opportunities for achieving STAR in a layered approach.

Layer 1. Methods for protecting receiver equipment from our own strong in-band signals are considered...

Layer 2. Ensure sub-systems also function without saturation so that behaviours are predictable and further noise (including harmonic and intermodulation signal products) is mitigated.

Layer 3. See underneath our own emission signals

Layer 4. Use prior knowledge of other signals in the EME in which we are operating in order to see signals of interest underneath them.

Commercial drivers for STAR (for example in internet Wi-Fi) consider energy and radio channel efficiency (radio channels are expensive to access). Metrics for such commercial applications may be measured in: J/Bit and Bit/s/Hz.

The primary military benefit is measured through an ability to continually sense anywhere in the EMS irrespective of our emissions.

How would a mathematical approach define this challenge in order to determine the limits of performance for STAR?

We expect the limits to be derived from combinations of factors such as:

- Nonlinearity in the transmitter and receiver circuitry
- Poor isolation between the transmitter and receiver resulting in coupling
- Thermal noise
- Modulation of our own signal
- The magnitude of signal to noise ratio we need to detect the other signal

Other Useful Reading:

A Widely Tunable Full Duplex Transceiver Combining Electrical Balance Isolation and Active Analog Cancellation – Laughlin 2015 - https://ieeexplore.ieee.org/abstract/document/7145660

Full-duplex Wireless: Design, Implementation and Characterization – Duarte 2012 - https://scholarship.rice.edu/bitstream/handle/1911/70233/DuarteM.pdf

Uncertainty in seismic inverse problems

Background

Seismic reflection data is used to estimate the properties of the Earth's subsurface from reflected seismic waves. The method is similar to sonar, echolocation, and medical imaging. One main task for a geophysicist is to convert the seismic response recorded on the surface to a 3D representation of the subsurface. Achieving the best possible quality of image and estimation of rock properties is critical in helping identify new reservoirs and understand the depletion and sweep mechanisms in existing reservoirs. Even marginal improvements here can significantly reduce the uncertainty associated with an opportunity, helping to improve our return on investment.

Inversion of multi-dimension of seismic data is typically achieved by a local optimization scheme, such as steepest gradient descent, during which the difference between modelled and observed seismic data (L2 norm) is minimized by iteratively updating the earth parameters (ie. Velocities, density, etc.). Such optimization method only yields a deterministic outcome and requires high data quality to succeed. To better quantify the uncertainty in earth parameter estimation, a probabilistic approach is needed. Bayesian techniques may be used to produce a posterior distribution that captures a range of possible earth models, thus providing a probabilistic outcome to the seismic inverse problem.

An MCMC algorithm

The Bayes theorem for our problem can be written as:

$$p(\boldsymbol{m}|\boldsymbol{d}^g) \propto p(\boldsymbol{m})p(\boldsymbol{d}^g|\boldsymbol{m})$$

where $\boldsymbol{m}$ is earth parameters sampled in depth, and $\boldsymbol{d}^g$ is seismic data in a gather sampled in time.

By assuming a Gaussian prior and observation errors, the prior and likelihood can be written as:

$$P(\boldsymbol{m}) = e^{\left[-\frac{1}{2}(\boldsymbol{m}-\mu_m^d)\Sigma_m^{-1}(\boldsymbol{m}-\mu_m^d)^T\right]}$$

$$P(\boldsymbol{d}|\boldsymbol{m}) = e^{\left[-\frac{1}{2}(\boldsymbol{d}^g-f(\boldsymbol{m}))\Sigma_d^{-1}(\boldsymbol{d}^g-f(\boldsymbol{m}))^T\right]}$$

where μ_m^d is the Gaussian prior mean, and $f(\boldsymbol{m})$ is the simulated seismic data. The operator f is a non-linear operator that acts on $\boldsymbol{m}$ and returns simulated data $\boldsymbol{d}_{sim} = f(\boldsymbol{m})$.

The posterior can be assessed by a random walk Metropolis-Hastings algorithm with the acceptance ratio

$$a = \min\left[\frac{p(\boldsymbol{m}_i|\boldsymbol{d}^g) \times q(\boldsymbol{m}_{i-1}|\boldsymbol{m}_i)}{p(\boldsymbol{m}_{i-1}|\boldsymbol{d}^g) \times q(\boldsymbol{m}_i|\boldsymbol{m}_{i-1})}, 1\right]$$

Due to the dimensionality of the problem, convergence of this MCMC algorithm is typically very slow. So we need to find ways to sample the posterior more efficiently.

Current solution

Making 'better' proposals

Often the proposal distributions are chosen to be symmetric (e.g. Gaussian, t-distribution). We can attempt to find an approximate posterior *P** that can be used as the proposal distribution. By simplifying the physics of the forward problem, consider the revised acceptance ratio as follows:

$$a = \min\left[\frac{p(\boldsymbol{m}_i|\boldsymbol{d}^g)p^*(\boldsymbol{m}_{i-1}|\boldsymbol{d}^s)}{p(\boldsymbol{m}_{i-1}|\boldsymbol{d}^g)p^*(\boldsymbol{m}_i|\boldsymbol{d}^s)}, 1\right]$$

where $\boldsymbol{d}^s$ is a stacked seismic trace rather than a gather of traces. To simulate $\boldsymbol{d}^s$ from the earth model $\boldsymbol{m}$, a linear operator f^* is used,

$$f^* = C \times D$$

where C is a seismic wavelet and D is a differential operator that converts earth parameters to a reflectivity series, and the simulated data $\boldsymbol{d}^s_{sim} = f^*(\boldsymbol{m})$. The approximate posterior $P^*(\boldsymbol{m}|\boldsymbol{d})$ is Gaussian,

$$P^*(\boldsymbol{m}|\boldsymbol{d}) = N\left(\boldsymbol{m}; \mu^*_{\boldsymbol{m}|\boldsymbol{d}}, \Sigma^*_{\boldsymbol{m}|\boldsymbol{d}}\right)$$

and the mean $\mu^*_{\boldsymbol{m}|\boldsymbol{d}}$ and covariance $\Sigma^*_{\boldsymbol{m}|\boldsymbol{d}}$ can be found by:

$$\Sigma_{\boldsymbol{d},\boldsymbol{m}} = f^*\Sigma_{\boldsymbol{m}}$$

$$\Sigma_{\boldsymbol{m},\boldsymbol{d}} = \Sigma^T_{\boldsymbol{d},\boldsymbol{m}}$$

$$\Sigma_{\boldsymbol{m}|\boldsymbol{d}} = \Sigma_{\boldsymbol{m}} - \Sigma_{\boldsymbol{m},\boldsymbol{d}}\Sigma_{\boldsymbol{d}}^{-1}\Sigma^T_{\boldsymbol{m},\boldsymbol{d}}$$

$$\mu_{\boldsymbol{m}|\boldsymbol{d}} = \mu_{\boldsymbol{m}}\Sigma_{\boldsymbol{d}}^{-1}\left(\boldsymbol{d}^s - f^*(\mu_{\boldsymbol{m}})\right)$$

where $\mu_{\boldsymbol{m}}$ and $\Sigma_{\boldsymbol{m}}$ are the prior mean and covariance, respectively.

Python code has been written for the above algorithm, but it has not been fully tested yet. There are some numerical issues encountered with the computation of the approximate posterior that need to be addressed. These numerical issues may be related to how the prior is specified, so prior model construction is also something that needs to be explored further.

Thoughts for some alternative solutions

Hamiltonian MC – Improves slow exploration of the parameter space by making proposals to the Metropolis-Hastings algorithm by considering Hamiltonian dynamics. The HMC makes use of the momentum variables to augment the target distribution. This introduces the computation of the kinetic energy associated with the state, in addition to the potential energy.

Reverse Jump MCMC - There are often many possible earth model, with different dimensions of the parameter space. Multi-model inference techniques like RJMCMC allows samples to be drawn from the posterior by jumping between different models. This is an attractive feature for seismic inversion because the subsurface structure is made up of layers of different thicknesses.

Hybrid MCMC – Combine the advantage of sampling efficiency of HMC with the desirable feature of multi-model inference offered by RJMCMC.

Approximate Posterior Inference – Variational methods. Posterior inherence is transformed into an optimization problem, where a variational distribution is introduced to approximate the actual posterior.

Note: datasets, algorithms and code are provided to the working group under a Confidentiality Agreement signed on behalf of the University of Cambridge.

PepsiCo Global Snacks R&D

Analysis of shear forces during mash disk formation

This problem relates to the production of chips in a particular process. A potato composite composed of multiple mashed ingredients and small particles is fed through a spreading manifold. This forces the mash into a cylindrical mould (forming process) on a rotating drum to produce mouth sized disk pieces, which then continue to a drying process.

During the forming process a scraper is used across the top of the mould to leave the mash flush with the top of the mould. This shearing process may affect the surface of the mash disk (as illustrated in the diagram below).

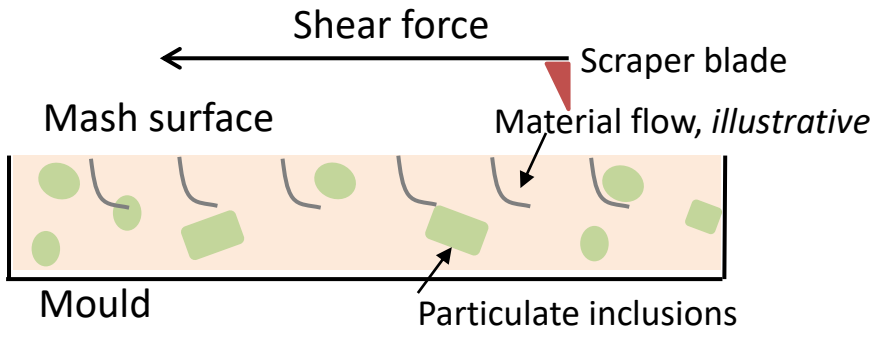
What is the effect of this shear on the material properties and structure of the mash disks and how can it be controlled? This will affect the texture of the final chip products, an important product attribute for consumers.

How would operating parameters and mould design affect the shear forces on the disk? This could impact scale up limitations and product design.

The key parameters are the composition of the mash, the pressure with which it is pushed into the former, the rotational speed of the drum, the shape / design of the scraper blade, force that pushes the scraper blade against the drum and the mould design. It is possible to attempt bulk rheology measurements on the mash.

Mathematical Challenge

Formulate a mathematical model of the scraping process and its effects on the surface of the disk and use this model to understand the dependence of the effects on the parameters that are available.



Towards managing landscapes: how can we interpret and design better environmental monitoring surveys?

Question

How should we interpret extensive environmental monitoring surveys and how can we design them better, especially in the light of different landscape factors, cropping and market share?

Background

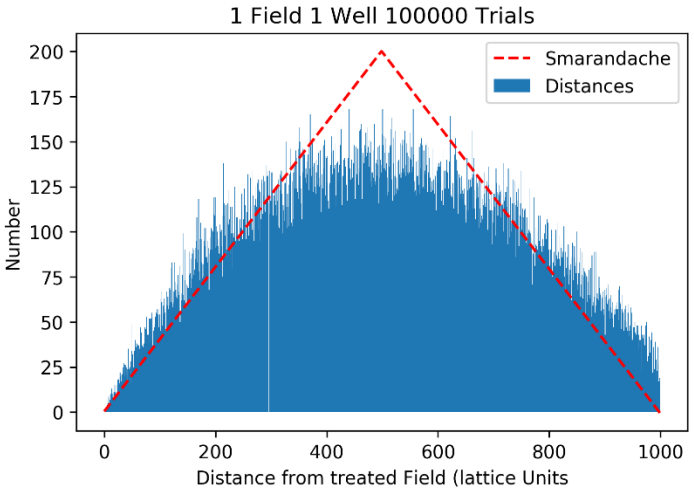
Member states routinely conduct groundwater monitoring surveys to demonstrate compliance of regulatory triggers for groundwater. However the link between the level of detect (concentration) and actual application is often not known, particularly when a large number of wells are monitored across a large landscape (say county level and above). It is therefore difficult to interpret the results of such monitoring exercises in the context of edge-of-field concentrations.

It is well-known that higher concentrations in groundwater are more likely to be found the closer a well is to an agricultural field. We would therefore like to know how varying the number of treated fields (combination of market share and cropping) and number of monitoring wells affects the probability distribution of wells close to treated fields.

An analysis where there are no restrictions on the placement of wells with multiple fields and a single well yields a $\sqrt{d}$ relationship, which would be intuitively expected. However the problem is more complex because wells cannot be placed anywhere: farmers do not like wells placed in fields because of difficulties with operating machinery and landscape factors such as roads and woodlands place restrictions on the placement of wells.

To simplify the problem we considered the random placement of fields on a lattice with the squares representing fields and the nodes possible well locations. Figure 1 shows 100,000 trials where a single well placed randomly in a grid and a centrally placed field yields a simple triangular distribution (Smarandache), however when randomly placed on the grid this relationship no longer holds (see Figure 1), and even with over 100,000 trials the exact form of the distribution is difficult to establish.

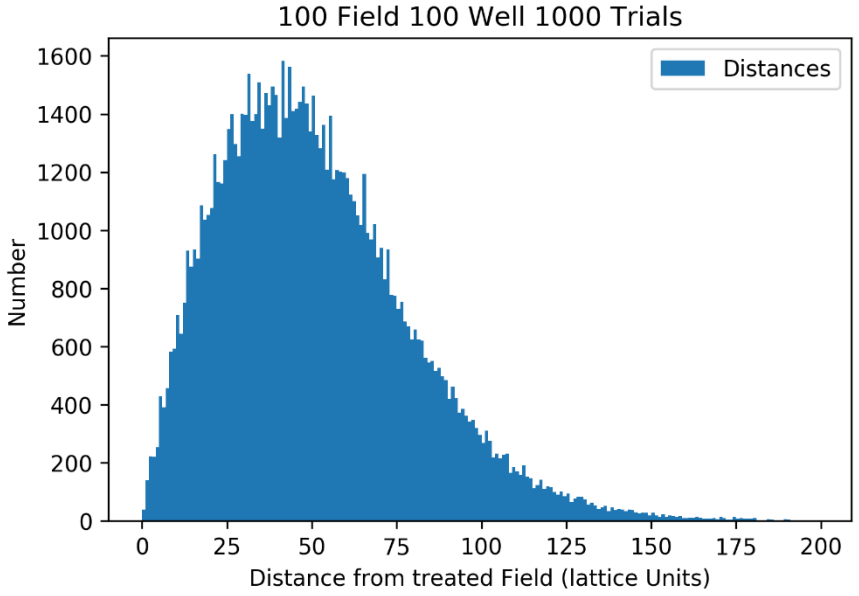
Figure1: relationship between distance to treated field with single well and field placed randomly on 1000x1000 grid



When the number of fields and wells is varied on a grid an approximate log-normal distribution is obtained (see Figure 2), but this relationship is not exact. We would first like to derive an exact relationship between the number of treated fields and proximity of monitoring locations on a lattice as this problem should be solvable and provide valuable insight into the real problem.

Real landscapes may be approximated by grids but are different: fields vary in size and shape, landscapes do not only contain agricultural fields, and groundwater may flow in a direction opposite to the monitoring well even if adjacent. We would like to know how different a real landscape is from the idealised problem and whether the results from the lattice can be bridged to landscapes. Ideally we would like to know how landscape factors affect how we should interpret and design monitoring surveys.

Figure2: Distance to treated field (100 wells and 100 fields on 1000x1000 grid) 1000 trials.



Subsidiary Questions

1. Are there any particular sorts of landscape arrangement that makes finding a well closer to fields more likely? E.g. lots of small fields versus fewer large fields
2. How does the probability change with the presence of non-agricultural areas such as towns, forests etc.?
3. Is it possible to take a general exceedance rate from a number of wells and predict an edge-of-field concentration?
4. Can any landscape be turned into an equivalent grid and analysed that way?

Identifying potential hardening techniques for image classifiers

Autonomous systems and machine learning is an ever accelerating field. In the specific case of object detection and image classification, research into neural networks and their applications rapidly became a main focus within the open source community. The famous ImageNet Large Scale Visual Recognition Challenge (ILSVRC) set the precedent in the field for developing networks with high levels of accuracy and as such, the security of these networks was often a secondary concern. These types of networks can be easily exploited using adversarial imagery. In which, perturbations are added to an image that will then cause the network to misclassify an image.

Often, these images are crafted to exploit specific types of classifier i.e. trained on certain types of data or use different architecture. Largely, this becomes an issue when the training of these types of classifiers is outsourced – how do we trust the system? Therefore, investigations into protecting these models are incredibly important. For this tasking, the possible different hardening techniques fall largely into two categories; data manipulation and network manipulation.

Data Manipulation:

- Can training data be manipulated sufficiently that a classifier trained on that data will be robust to a variety of adversarial methods?
- Can extra, statistical information be extracted about the data and fed in at the training stage (i.e. extra information that an adversary cannot access).

Network Manipulation:

- Can hardening be introduced into the pipeline before the training stages? I.e. can the architecture of the network be altered such that adversarial images introduced in training have little to no effect on the network?
- Is there a way to tell if a network has been compromised by bad data after it has been trained? E.g. “badnets” [1].

It is likely that combinations of the above will be required to harden a neural network sufficiently against the majority of attacks. However, insight into why individual components of the system i.e. data, architecture, functions have certain effects on the behaviour of the system would be extremely beneficial.

References:

[1] https://arxiv.org/abs/1708.06733

Notes