# Potential project ideas

Jonas Lindblad

**Abstract**

This document outlines some of the project ideas I have been thinking about for the Bayesian Data Analysis course. Excuse my brevity, as I have not had a lot of time to think/research.

## Contents

## Idea 1: Meteorological datasets

I have previously worked with meteorological data quite a bit so I have some knowledge of common data formats (grib files, netcdf, etc) and sources (US government, ECMWF). I think meteorological data would be a great source for this project. We only need to make some decisions on what type of variables and model to include.

### Data

#### Numerical weather prediction datasets

One key potential source is the European Centre for Medium-range Weather Forecasting (ECMWF), as they provide certain restriction-free datasets: ECMWF Public Datasets. Among these perhaps the most important one is the ERA5 reanalysis dataset. This is a gigantic set of variables over the entire Earth; the data is in a format ready for input into the numerical weather simulations they run on their supercomputer cluster in Switzerland. Spatially, the data is organized in a spherical grid, with height levels given by pressure levels (reminder: pressure is higher at surface level, so the height variable actually becomes smaller as you move higher up in the atmosphere). Temporally, the data is based on 12-hourly empirical observations and hourly simulation steps in-between.

It could be useful to click around on the ECMWF website and explore what variables could be used. We need to decide on one observation variable $y$ (this can also be a forecast) and some covariates $X$. But I expect that this can be changed any time or after some data exploration.

There are also many other public datasets of similar type, but some may be difficult to use. The ECMWF datasets are in Grib format, which should be readable with Python. If it's not too difficult, we should probably process the data in Python and later use R for the main analysis. Another data source which is similar to the ECMWF data is the US government numerical weather models. I guess the main one to look at would be the GFS model, link: GFS model data. This data is similar to the ECMWF data, but they have empirical observation ever 6 hours instead of 12-hourly. There are also other models, but I know much less about them. I think the ECMWF model has the advantage of more available variables, but not sure if they are anything interesting to model.

Nevertheless, these are all global models and based on rather sparse grids. One interesting question may be to model the error of these weather forecasts by comparing with measurements from weather stations.

### Weather observation stations

Another obvious data source is the direct empirical observations from weather stations. In Finland the important source would be the Finnish Meteorological Institue (link here: Download FMI observations). It's well worth exploring the FMI website. One thing to note is that the data is provided in CSV table format, which makes it exceptionally easy to use (especially compared to the global gridded data for numerical weather prediction).

I think other weather station data should also be available as most governments tend to release scientific data with no restrictions for public use. But I have not looked into it much.

## Models to use

It's not clear to me what kind of model should be used, but one obvious idea is some kind of time series modeling. Generally we should choose one outcome variable (e.g. daily precipitation) and let other data be covariates. Then we formulate some kind of model with unobserved quantities in some suitable way. We could always read some basic meteorology to get ideas.

# Idea 2: Hurricane forecasting

This is the only project where there is a basically ready-to-use dataset. So, in hurricane forecasting there are usually two quantities of interest - track, and intensity. We can focus on either of them but I think intensity would be easier. The US government of course uses multiple models for hurricane forecasting; you would usually imagine very computationally heavy numerical models that simulate the laws of physics. There is some of that, but one of the best intensity forecasting models is actually a multivariate linear regression (really!). This model is called Statistical Hurricane Intensity Prediction Scheme (SHIPS). We don't need to think too much about how they compute the variables for SHIPS, but it's important to keep in mind that it's a large model with over 100 covariates.

## Data

You should have a look at the SHIPS website, link: SHIPS.

From the SHIPS website you can find all the relevant information and scientific articles related to SHIPS. More importantly, under 'developmental data' you can find links for '5-day SHIPS predictor files', these files would act as our dataset. They are initially in a kind of difficult form, so a first step in the project could be to process them into CSV format that is easier to deal with.

In case you are interested, we can also look for hurricane track datasets. The main datasets of interest here would be HURDAT2 and IbTracs, I think. I am not very familiar with these but I expect that we could think of some Bayesian analysis to do with them.

## Models to use

The SHIPS data should be laid out so that it's easy to predict the future intensity (VMAX) for every 6th hour (+6h, +12h, +18h, . . . ). I would suggest we pick either 24h or 48h to begin with. Then we could pick out some subset of covariates (I can probably find the most important ones) and use a linear regression model. We should probably not think too much about priors, and just perform some MCMC simulations on this kind of model. I think this would be a great first step in the project and I expect that everything I've described so far is doable without too much extra studying.

Note that on the SHIPS website there is data for the different basins: Atlantic, Pacific, Southern Hemisphere, Indian Ocean. One idea to try would be to design a hierarchical model with respect to the basins (I am not sure if this would work, but it's an idea to try).

## Idea 3: Bayesian modelling of Google trends keywords

In case this is unfamiliar to you, Google has a service that lets you ask for number of Google searches over time for words and phrases. You can explore the service with this link: Google Trends

The basic idea for the project (I think, but we can discuss this) would be to do some kind of time series modelling that could potentially be used as a component of trend forecasting. I do not have a 100% clear idea of how this project should be done, and unfortunately this project may be a bit more open-ended than the other ideas. It's also possible that Google Trends could be used in some way in combination with some other data/idea (some kind of difference-in-difference model?).

### Data

An unfortunate limitation is that the data is always returned pre-scaled to the (0,100) range and rounded to nearest integers. Since you can supply multiple words it is still possible to do some kind of multiple-search comparison, but one needs to keep in mind that if one search word is on average less frequent than another, you will lose some resolution due to the scaling and rounding. I do not know if the 'raw' data is available somewhere, but this may be worth checking. Another potential issue is the limited time resolution; it seems that for 90-day ranges you can retrieve one data point for every day, but for >1 year you will get only one data point per week.

Below are some example queries, where I search for data on the keywords 'Mathematics', 'Physics', and 'Chemistry'. The first search is for a 90-day range, while the second is a 5-year range. Note that in the former search there is a weekly pattern of lower search numbers during the weekends. In the latter search this pattern disappears due to the worse time resolution.

https://trends.google.com/trends/explore?date=today%203-m&geo=US&q=Mathematics,Physics,Chemistry

https://trends.google.com/trends/explore?date=today%205-y&geo=US&q=Mathematics,Physics,Chemistry

Of course, these queries are just an example. If you can think of anything that would be fun to research, we can discuss things and try out different words in Google trends.

### Models to use

I am not very familiar with time series modeling or with how Bayesian data analysis is usually done for time series. For that reason I would ideally want to do this project with a partner who has taken one or two courses about time series analysis. That aside, I do not really expect it to be difficult to learn as we go, if this project seems especially interesting to you.

Another possibility would be to generate data for a difference-in-difference model (I try to explain this kind of model, but it would probably be best to Google a bit. I found this website which explains it well: Health Policy Data Science). The idea in a DID model is that you have some set of values $y = (y_1, \ldots, y_N)$ that you can observe at two times - before and after some effect (treatment/experiment). In other words you have two vectors of observations $y(1), y(2)$. The $y$-values are expected to change based on some associated variables $x_1, \ldots, x_N$ and usually there is also some control variable (e.g. $y_1$ could be control) which is expected to act as some kind of 'baseline'. In any case, the idea is to use Google trends data and write down a Bayesian model for

$$(y_i(2) - y_j(2)) - (y_i(1) - y_j(1)),\ i \neq j,$$

so we actually model how the difference between $y_i$ and $y_j$ changes after the treatment/effect. But we can discuss the specifics later. The important thing to understand is just that Google trends could be used to observe changes in search terms before and after an event, and to get separate data for different countries/regions. One thing that immediately comes to mind is the current US election, but it may be too much of a hot topic and not suitable for a course like this. But I expect it could be quite fun to look for patterns related to this in the search data.

# Idea 4: the cube-root law in voting theory

In voting theory the researcher Taagepera formulated a kind of 'formula' for the national assembly size as a function of country population. Obviously larger nations will have larger assemblies. But did you know that the assembly size $S$ of a country tends to be approximately

$$S = P^{1/3},$$

where $P$ is the country population? The basic idea here would be to model the assembly size in some suitable way with a Bayesian model. The most logical way of doing this would probably be to look up official data for a large number of countries, and create a data set and do some inference on the parameter $\alpha$ in

$$\log S = \alpha \log P.$$

Originally Taagepera wrote a paper giving some theoretical justification for why $\alpha \approx 1/3$ **should** hold. In the simplest case, we can just let $\alpha$ have a (somewhat) strong prior with mean $\mu_0 = 1/3$ to reflect this. Then, using the data we can compute a posterior interval for $\alpha$.

However, this is not much of an analysis. I think more complex models should also be considered. For example, population increases exponentially but the assemblies can only be expanded by legislation. The US congress has not been expanded since the early 20th century. This suggests that the rate of population growth $g$ and the time since last expansion $T_e$ must be considered as well. To express this mathematically,

$$\log S = \alpha \log P_e, \quad P_e = Pe^{-gT_e},$$

here $P_e$ means 'population at time of assembly expansion'. To understand that $P_e$ is defined correctly, just multiply it with the population growth factor $e^{gT_e}$. In any case, this allows us to write a better model

$$\log S = \alpha \log P - \alpha g T_e.$$

Probably this model has some trouble as well, but it's still worth doing some analysis with it. One problem is that you may want the exponential growth factor changed to one with varying growth rate $g = g(t)$. Then instead of $gT_e$ we would write $\int_0^{T_e} g(t)dt$. This model may not be suitable for a Bayesian analysis, however. Another objection might be that human populations don't actually grow exponentially but instead follow a logistic equation (see link: Logistic equation for population).

Another question is if economic variables could have some interaction with the $\alpha$ parameter. One could imagine that in a wealthy country there would be a larger number of politicians per capita. Another interesting question is interaction with literacy rates and maybe rates of higher education. But I have not really looked into this much.

## Data

Main data source could be the World Bank, maybe? At a first glance it's not obvious but common sense would say that population/economic data can't be difficult to find.