Luminosity

Cameron, M. Rosales, V. Westermann, J.P.

July 30, 2017

Contents

List of Figures

List of Tables

Notes (Please Read):

- print(df.to_latex()) will give a latex export of the dataframe in the jupyter notebook, you can use this to quickly copy paste into latex
- plt.savefig(<path>) will allow you to save your figure as a png so it can be used in the document/presentation
- Make sure to commit and pull as much as possible to avoid merge errors
- Don't edit the styles of this document yet, will do that at the end

1 Introduction

1.1 Summary

TODO Will do this last

1.2 Literature Review

1.2.1 Luminosity-based Approach

TODO Michael: Try to accumulate as much as possible. We have such a long list of papers anyway.

1.2.2 Natural Disaster Economics

TODO Viviana: obviously, as the expert.

2 Data

2.1 Data Description

TODO Micheal: Describe what the data looks like, how many observations there are, where we got it, who else has used it etc.

2.2 Data Preprocessing

Volume While the number of observations is extremely low, as only annual images are available from 1992 to 2013, the dimensionality of each observation (image) is considerable. With a size of 16801 by 43201, every image contains 725820001 pixels in total, which results in more than 700MB of disk-space required for only one image in uncompressed format. This also means that computations on the entire dataset are not possible with common personal computing architecture.

Diverse Sources To aggregate meaningful information to use in the model, such as economic indicators from the world bank (gdp, import, export, inflation, etc.), we need to join data from a lot of different sources on different aggregation levels.

Image Format As the data comes in image format, we need to find an appropriate numerical representation of the images.

2.2.1 QGIS

TODO Viviana

2.2.2 Python Architecture

To enable exploration and flexible modelling of the data even on a platform where memory and computing power are quite restricted, we built a supporting architecture in python for loading the images and aggregating the features. To quickly load entire time series of the most important places, the sections of the image representing the biggest cities and a 300x300 area around them are cut out and stored as seperate time series. This allows for more easilty training predictive models for the luminosity. Additionally, functions where created to easily load a 'section' time series (the series of images for one particular subsection of the world's image), given coordinates and the desired size of the image. The stored data is no longer in the *tif* format provided at download. All images have already been passed into arrays and stored as compressed numpy arrays. That reduces the required storage space by an order of magnitude. The functions to preprocess and load the data can be found in the GitHub repository. To make sure that the code doesn't throw a *MemoryError* or similar when working on a less powerful machine, most of these functions have been optimized to free up memory space

as fast as possible and always only work with one single memory mapped full-size image array.

3 Modelling

3.1 Disaster Impact Models

Exploratory analysis and some basic models are a first step to assessing the impact of natural disasters on the luminosity time series. For this, we need to make a modelling decision regarding how the disaster (represented only as a single point location) can be geospatially associated with pixel luminosity values. In the following we will quickly introduce and discuss two different approaches to this proble, one that is less conservative and one that complies with most of the research currently published in the area.

3.1.1 Linear Distance-based Modelling

One practical mathematical choice is a function decaying with distance affecting areas or even individual pixels on the grid. The advantage of this method is that it is simple to explain, easily tuneable and leaves a lot of flexibility for modelling. Additionally, it captures the notion that areas in the vicinity of a disaster event are more likely to be affected than those further away by default. We trialed this approach using the luminosity data in combination with location-tagged earthquakes from the U.S. Geological Service. We extracted 150x150 squares from the satellite image of the entire planet in a similar fashion as a convolutional neural network would apply to an image. Then we filtered the top $t\%$ most luminous 'subimages', as we can generally disregard those parts of the world that contain no lights (oceans, deserts, etc.) and calculate a disaster coefficient based on a function decaying linearly with distance. The t percentage threshold can be tuned in order to keep the data small enough to process. This is necessary as in practice, finding the disaster coefficient means calculating the distance between every image section's center and every earthquake and then applying a linear function of the distance and the magnitude for every earthquake that occurs in a given year, then summing up these effects to generate a disaster impact variable for that observation. This results in a high computational load as for every section that we extract from the image, we need to compute the eucledian distance to every singe one of the 25000 earthquakes, making in

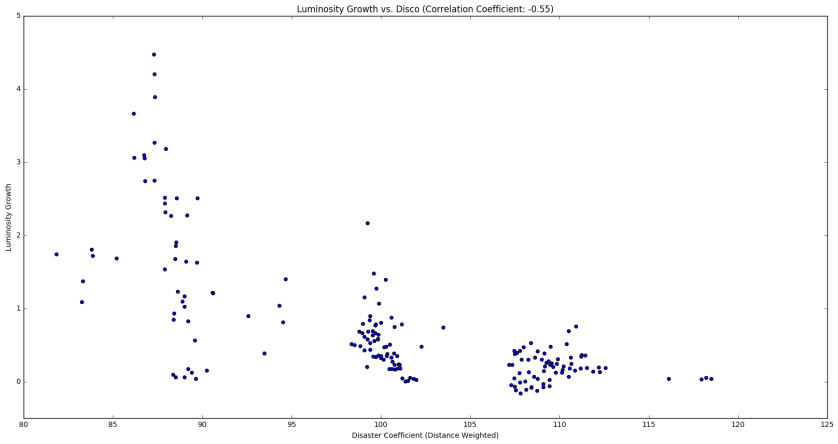

Figure 1: Luminosity Growth 1992-2013 plotted against a linearly decaying disaster coefficient for 150x150 image sections.

almost infeasible on a common laptop. However, already only the luminosity growth between beginning and end of the image time series and the disaster coefficient show promising qualities of the data when plotted agains one another. As we can see, there is a negative relationship between (luminosity) growth and the impact of earthquakes. The effect has been aggregated over the entire series and we can thus look at this as just showing the general implications for growth of a city/region by being in or closer to an area that has more earthquakes. It is notable that not necessarily the actual growth value, but the variance of the growth values strongly increases for areas less prone to earthquake effects.
However, this approach makes some strong assumptions about the nature of natural disasters that don't hold in reality. An important factor in how much impact a disaster has on a region are geographical features: An earthquake will affect different areas differently based on their rockbed and geological consistency while e.g. storms and floods depend strongly on the topography. Nevertheless, the calculation of such a disaster coefficient yields promising results and suggests there is a relationship in the data between disaster coefficient and the growth in a location, proxied by luminosity.

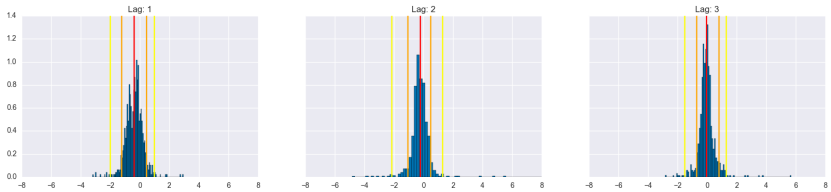

Figure 2: Distribution of Lag Coefficients for Earthquakes in Vector Autoregression Models per City with 95th and 99th Percentiles

3.1.2 Recorded Disaster Impact-based Modelling

Another possible approach, one that is far more commonly applied in Economics, is to use institutional data on which cities where affected by an earthquake. Using the EM-DAT international disaster database, we can obtain records of cities that were affected by a given earthquake in a given year. Matching this with a simple list of the world's largest 50000 cities, we can obtain a panel dataset that associates earthquake damage without any measure of magnitude to cities, some affected, others not. Using this dataset and calculating luminosity sums for each of the areas around a city, we can fit a simple vector autoregression model to explore the impact of earthquakes on these time series. The model is enriched with economic indicator values as well as import and export statistics and fit for every city indipendently. To get a general overview of the relationship between luminosity and the disaster events, we can simply average the earthquake lag coefficients for all models. Doing this yields the following average coefficients (the distribution of which can be viewed in Figure TODO.

	Average Coefficient
el1	-0.36
el2	-0.21
el3	-0.03

These results support those found using the alternate approach, that there is a measurable effect of earthquakes on the summed luminosity for an area. As the distribution of the lag coefficients suggest, the impact is far stronger in

the first two years after the earthquake which is when things normalize. As these coefficients are aggregates, however, they do not provide much information on the significance in a specific case or the detailed recovery process. Nevertheless, these preliminary results in combination with those collected using a distance-decay impact model, are important baselines for a more elaborate panel model.
While being based on fewer assumptions about the nature of a distaster's impact radius and magnitude due to using human assessment-based data in every case, this approach has an obvious shortcoming the alternative did not: As we do not imply any kind of decay with distance from the earthquake location or use information on the earthquake location and magnitude at all, we are entirely neglecting the intuitive notion that the impact of the earthquake differs from place to place. Because of these shortcomings, future research should utilise a combination of the two approaches, using also distance from the earthquake for any given city as part of the function that describes disaster impact.

3.2 Panel Model

3.2.1 Region-based Panel

TODO Viviana

3.2.2 Section-based Panel

Using the list of cities affected by earthquakes created from the EM-DAT disaster database, we can construct a panel dataset containing all years available (1992-2013) and a selection of the world's most populous cities. To achieve this, we simply take a 50x50 pixel square around the city's coordinates to calculate the luminosity. This can be combined with the other data described in the previous section to form a more granular panel that can study effects on a city level, rather than the very aggregated regional level. The advantage of this method is also that time-series constructed from city locations are more homogenous than those constructed on regional level. For an overview of what the time series used with this methodology look like in terms of the actual image who's luminosity is summed up, you can refer to Figure TODO.
A possible extension to this methodology that is beyond the scope of this paper, would be to extract the actual individual settlements, thus ignoring

surrounding locations that are simply included because of the frame size and shape.

3.2.3 Dynamic Panel

TODO Viviana: Describe here how the model that you are using is constructed, where you got it, etc.

4 Results

4.1 Case Analysis

TODO Micheal: This is where your case analysis for different places goes, try to add some statistical tests etc. if possible. E.g. distribution of light one year vs the next compared to overall time series distribution changes (shocks).

4.2 Modelling Results

TODO Viviana: Describe the results of the regression here, significant values and what those values mean.

4.3 Conclusions

TODO Will do this just before the summary

4.4 Outlook

TODO Jonas

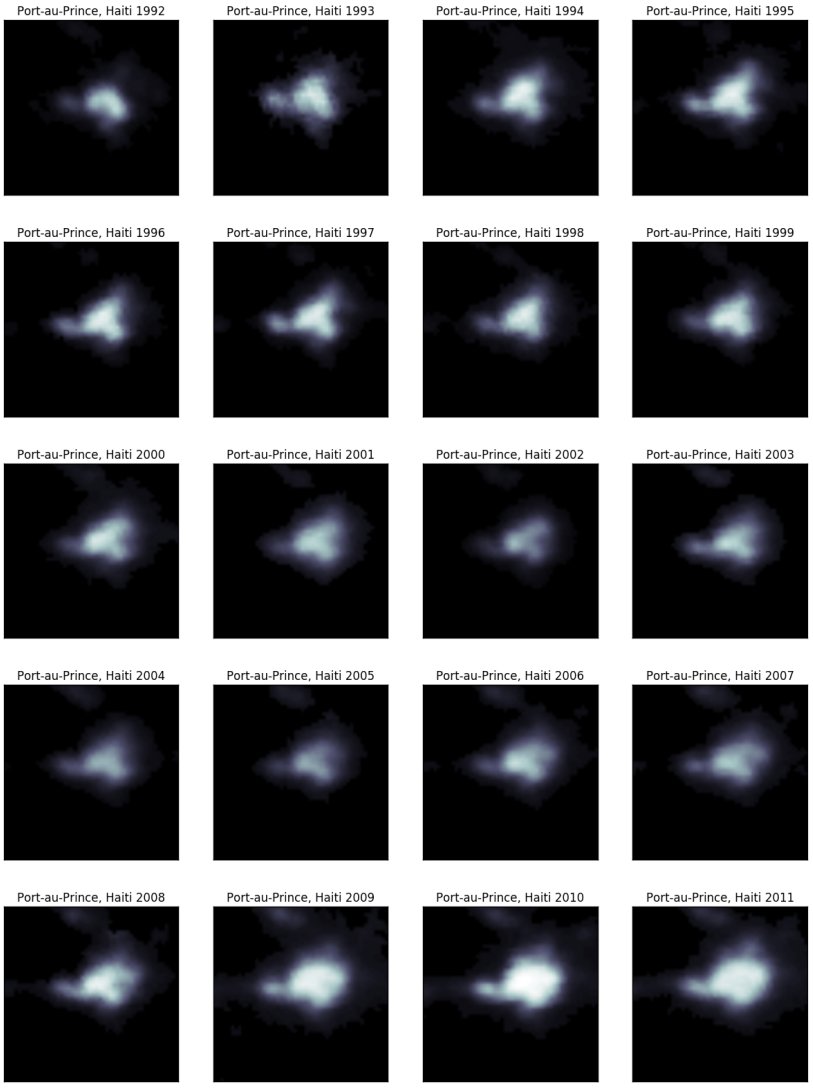

Figure 3: 50x50 pixel atellite image cutout of Port-au-Prince, Haiti