Comparison of SVD and LSTM using time series data of premature deaths due to air pollution from the Balkans

Sophie Logan
University of Chicago
Chicago, IL
slogan@uchicago.edu

Rodrigo Salazar
University of Chicago
Chicago, IL
rodrigosalazar@uchicago.edu

1 Introduction

Despite the incredible technological innovations around the world within the last 50 years, deadly problems like premature deaths caused by air pollution - for which resources to solve the problem exist - remain persistent issues for economic and political reasons. In the past 30 years, many economies have made great progress mitigating this issue, especially in North America and Western Europe. Unfortunately, other regions have not made such significant advancements.

One of these regions is the Balkans, the area that, historically, has had the highest mortality rates by air pollution in the world. Dangerous levels of particulate matter, according to WHO standards, were surpassed for 120-180 days in a year, whereas the EU limits those levels of particulate matter to only exceed 35 days/year. In the countries of this region most of the particulate matter comes from household heating and thermal power plants using low quality lignite coal. Also, they suffer from energy poverty, meaning that households spend a large percentage of their incomes on energy. To combat this pollution issue, policymakers are calling for energy equity measures to make it more feasible for households to transition to cleaner electricity sources, however, this has not been enough and mortality rates are still very high. In addition to the toll on health, costs due to premature deaths amounted to over 13% of their annual GDP in 2019.

Regions comprised of small countries, like the Balkans, share many characteristics economically, politically, socially, and environmentally. The connections between countries is especially salient for an issue like air pollution, where the particulate matter from one country influences the pollution in another. As a result, the mortality rates within this region are correlated with each other. For this reason, we want to analyze the historical mortality rates caused by air pollution from the past 30 years in the countries of this region in order to fit a model that can forecast the evolutions of these rates. To do this, we propose two approaches, one is using SVD1, which has been used successfully to forecast mortality data in some countries, and the other is an LSTM-autoencoder which has been more used recently 4 because it is capable of modeling better non-linear relationships in time-series data.

2 Data

We are using the *Mortality, morbidity and welfare cost from exposure to environmental-related risk* database from the Organization for Economic, Co-operation and Development (OECD). Initially we filter the data to obtain the premature deaths per million people related to ambient particulate matter exposure from 1990 to 2019 of more than 200 countries. We analyzed the data and found regional patterns, where we realized the alarming rates in the Balkans region and decided to focus our work there. Thus, we filter our data again to only consider Albania, Bosnia and Herzegovina, Bulgaria, Croatia, Montenegro, North Macedonia, Romania, and Serbia.

3 Literature Review

In 1992, Lee and Carter, successfully forecasted US Mortality using SVD. Their paper became a reference for using SVD with model demographic data. Their X matrix is composed of the mortality rates $m_{x,t}$, where the rows are the years and the columns the age groups. Then, they proposed they seek the least squares solution to the model:

$$ln(m_{x,t}) = a_x + b_x k_t + e_t \tag{1}$$

Where a_x is the mean age-specific mortality schedule across all years of analysis, b_x is the average contribution of age group x to overall mortality change over the period, and k_t is the incremental change in period t. We use a similar approach, but instead of age groups we use the Balkan countries. We have seen that the Lee model was recently extended using an autoencoder (Miyata et. al., 2022).

A basic autoencoder requires two functions - an encoder and decoder - that can reduce data to its most significant variables. For the LSTM-autoencoder model we use as our main reference the work of Xayasouk et. al. 2020, where they used long short-term memory (LSTM) and deep autoencoders (DAE) to forecast pollution in South Korea based on hourly data from 25 measuring stations around Seoul from 2015-2018. First they construct an LSTM model. The key to an LSTM model are the input, forget and output gates used to create the hidden layer. Each "gate" uses weights and a classification function to "decide" which choice of weights passes to the next stage. They then construct a deep autoencoder, which is essentially stacked autoencoders.

Their LSTM model performed slightly better than the DAE model and was able to predict the particulate matter for a particular area where a user is located. We modify the setup of their problem to work with countries rather than regions within an urban area and have annual pollution data over 25 years rather than hourly data for 3 years. We also adjusted the model to be an LSTM, hoping to get the benefits of both of the models that they proposed.

4 Models

As we mentioned earlier, we build our mortality matrix $m_{x,t}$, using countries in the columns and years in the rows.

4.1 SVD

Using the SVD model we mentioned earlier Lee and Carter used:

$$ln(m_{x,t}) = a_x + b_x k_t + e_t \tag{2}$$

Where a_x is the mean country mortality across all years, b_x is the average contribution of country x to overall mortality change over the period, and k_t is the incremental change in period t. To estimate ax we simply calculated the means of the log(mortalities) for each country. Then we substract that from the data resulting in $ln(m_{x,t}) - a_x$. By decomposing this matrix into the SVD form:

$$U\Sigma V^T \tag{3}$$

We estimate b_x as the first right singular vector, and k_t as the first left singular vector times the first singular value. We only use data to get our estimates for the first 20 years, so we can do the forecast by estimating the k_t values using an ARIMA model and get the mortalities for the next 10 years and compare them with the true values.

4.2 LSTM

LSTM is able to store both short-term and long-term "memory", which makes it good for analyzing sequential data such as time-series. To do this, the LSTM uses three mechanisms: an input gate, a forget gate, and an output gate. The forget gate functions to choose whether to keep the current hidden state output or resort back to the output from the previous hidden state. It does this by multiplying both layers by the weights, adding the bias, and putting the entire expression through a sigmoid function in order to get a value between 0 and 1 where the forget gate can decide which layer to move forward.

The input gate is used to transfer information from the forget gate to the cell state. It takes in the hidden state and x at time t and runs them through a sigmoid function to create a vector with values between -1 and 1.

The output gate takes in the cell state and decides which information to output. This happens in three steps:

- A vector is initialized and the hyperbolic tangent function $\tanh$ is applied to the cell state to scale the values from 1 to 1.
- The sigmoid function is then applied to the previous hidden state to create a filter for values of h_{t-1} and x_t.
- Lastly, the filtered values are multiplied by the vector created in step 1 to produce LSTM output information.

For each element in the input sequence, each layer computes the following functions:

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \tag{4}$$

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \tag{5}$$

$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \tag{6}$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \tag{7}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot g_t \tag{8}$$

$$h_t = o_t \cdot \tanh(c_t) \tag{9}$$

Where W is our matrix of weights, x is our input data, b is our bias, h is the hidden state, c is the cell (and acts as the memory), x is the input and i, f, g, and o are the input, forget, cell, and output gates respectively. In our multi-layer LSTM, the input for each layer is the previous hidden state of the previous layer multiplied by a dropout rate, which is a Bernoulli random variable.

We use the MSE as our loss function:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i, \tilde{y}_i)^2 \tag{10}$$

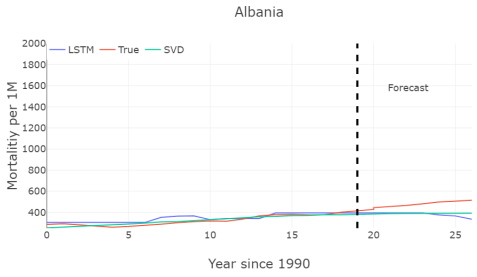
5 Results

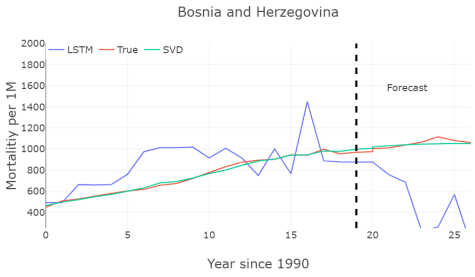
We tried a variety of hyperparameters to try to fit our data, especially for non-linear functions like the case of Bulgaria. We experimented with learning rate, sequence length, and number of hidden layers to minimize the errors. After running the SVD and the LSTM models, we found that for half of the countries one model does better than the other.

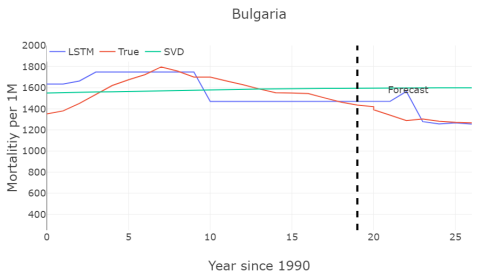
After running both SVD and LSTM to forecast the last 10 years of our data we got the next results:

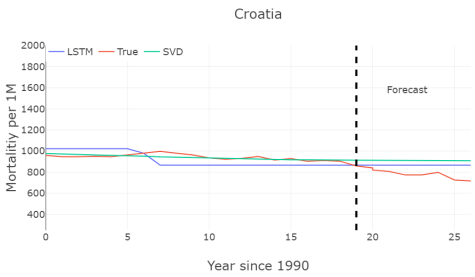
RMSE	Alb.	Bos. Her.	Bul.	Cro.	Mont.	N. Mac.	Rom.	Serb.
SVD	79.42	26.9	252.9	119.7	47.21	62.01	105.27	165.73
LTSM	93.1	533.33	101.78	83.87	66.86	14.44	43.86	200.68

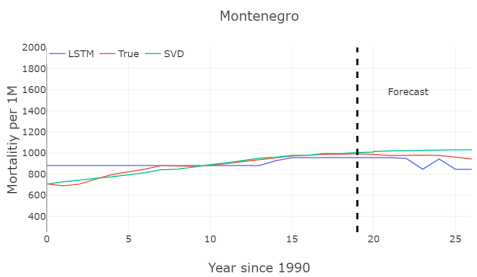
The errors and the graphs show how both models have a good performance to estimate the mortality rates. Although, we were expecting the LSTM to perform better due to its capacity to capture non-linear relationships in the data. These results show us the importance of revisiting cases where we were using models like SVD and trying more recent techniques that could potentially perform better.

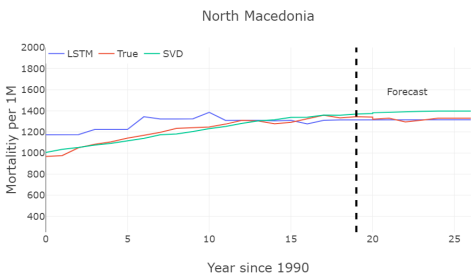

Albania
LSTM
True
SVD
Forecast
Mortalitiy per 1M
Year since 1990

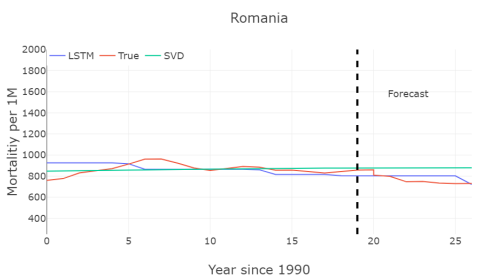

Bosnia and Herzegovina
LSTM
True
SVD
Forecast
Mortalitiy per 1M
Year since 1990

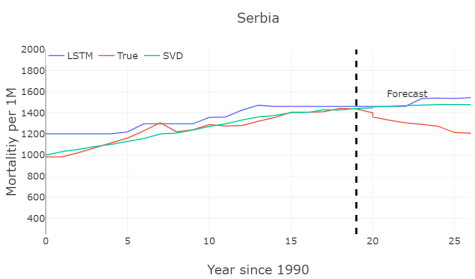

Bulgaria
LSTM
True
SVD
Forecast
Mortalitiy per 1M
Year since 1990


Croatia
LSTM
True
SVD
Forecast
Mortalitiy per 1M
Year since 1990


Montenegro
LSTM
True
SVD
Forecast
Mortalitiy per 1M
Year since 1990


North Macedonia
LSTM
True
SVD
Forecast
Mortalitiy per 1M
Year since 1990


Romania
LSTM
True
SVD
Forecast
Mortalitiy per 1M
Year since 1990


Serbia
LSTM
True
SVD
Forecast
Mortalitiy per 1M
Year since 1990

References

[1] Franco, Edian F., Pratip Rana, Aline Cruz, Víctor V. Calderón, Vasco Azevedo, Rommel T. J. Ramos, and Preetam Ghosh. 2021. "Performance Comparison of Deep Learning Autoencoders for Cancer Subtype Detection Using Multi-Omics Data" Cancers 13, no. 9: 2013. https://doi.org/10.3390/cancers13092013

[2] Lee, Ronald D., and Lawrence R. Carter. “Modeling and Forecasting U. S. Mortality.” Journal of the American Statistical Association 87, no. 419 (1992): 659–71. https://doi.org/10.2307/2290201

[3] Miyata, Akihiro, and Naoki Matsuyama. "Extending The Lee–Carter Model With Variational Autoencoder: A Fusion Of Neural Network And Bayesian Approach." ASTIN Bulletin: The Journal of the IAA 52.3 (2022): 789-812.

[4] Xayasouk, Thanongsak, HwaMin Lee, and Giyeol Lee. "Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models." Sustainability 12.6 (2020): 2570.