Zurich University of Applied Sciences

School of Management and Law

Master of Science in Banking and Finance

Deep Learning

Assignment

-

“Air Quality Prediction”

Professor:

Dr. Bledar Fazlija

Department of Quantitative Finance

Submitted by:

Andreas Bittel (16-605-289)

Severin Marchetti (15-535-222)

Mark Prenrecaj (17-474-784)

Fabio Tanchis (16-570-210)

June 20th, 2022

Statement of truth

"We hereby declare that we have written this thesis independently, without the assistance of third parties and using only the sources indicated, and that we will not hand out copies of this thesis to third parties without the written consent of the course director."

At the same time, all rights to the work are assigned to the Zurich University of Applied Sciences (ZHAW). The right to citation of authorship remains unaffected.

Student name

Andreas Bittel

....

Signature of student

Student name

Severin Marchetti

....

Signature of student

Student name

Mark Prenrecaj

....

Signature of student

Student name

Fabio Tanchis

....

Signature of student

Table of contents

Statement of truth I

Table of contents II

List of figures III

List of tables III

List of abbreviations IV

1 Abstract 1

2 Introduction 1

3 Theoretical Framework 2

3.1 OLS Regression 2

3.2 Recurrent Neural Networks 4

3.3 Mean Squared Error 6

4 Literature Review 7

5 Results 7

5.1 OLS Regression 8

5.2 LSTM 8

5.3 Threshold 10

6 Conclusion 10

7 Bibliography XI

8 Appendix XIII

8.1 Github XIII

8.2 OLS Results XIII

8.3 LSTM results with random parameters XIV

8.4 LSTM results with tuned parameters XVI

8.5 Threshold results XVII

List of figures

Figure 1: Simple linear regression (own representation based on Bachmann, 2020, p. 8 ff.) 2

Figure 2: Simple linear Regression with OLS (own representation based on Bachmann, 2020, p. 16) 3

Figure 3: an unrolled recurrent neural network (based on Cola, 2015) 5

Figure 4: Structure of a LSTM cell and equations (Varsamopoulos et al., 2018, p. 4) 5

Figure 5: OLS for Florida (own representation) 8

Figure 6: LSTM with tuned parameters for Florida (own representation) 9

Figure 7: OLS for Arizona (own representation) XIII

Figure 8: OLS for Kansas (own representation XIV

Figure 9:LSTM with random parameters for Arizona (own representation) XIV

Figure 10: LSTM with random parameters for Florida (own representation) XV

Figure 11: LSTM with random parameters for Kansas (own representation) XV

Figure 12: LSTM with tuned parameters for Arizona (own representation) XVI

Figure 13: LSTM with tuned parameters for Kansas (own representation) XVI

Figure 14: Air quality results for Arizona (own representation) XVII

Figure 15: Air quality results for Florida (own representation) XVII

Figure 16: Air quality results for Kansas (own representation) XVIII

List of tables

Table 1: Tuned parameters (own representation) 9

Table 2: MSE values for random and tuned LSTM models (own representation) 9

List of abbreviations

CO	Carbon Monoxide
et al.	et alia
etc.	et cetera
GRU	Gated Recurrent Unit
LSTM	Long-Short Term Memory
MSE	Mean Squared Error
NO2	Nitrogen Dioxide
O3	Ozone
OLS	Ordinary Least Squared
RNN	Recurrent Neural Networks
SO2	Sulfur Dioxide
ZHAW	Zurich University of Applied Sciences

1 Abstract

In this paper, the air pollution problem is addressed in more detail by attempting to predict the CO values within the atmosphere using two different modelling approaches. Specifically, this was done using the multiple linear regression's OLS and the RNN's LSTM approaches. The performance of these approaches is measured with the MSE. Furthermore, thresholds are based on quartiles to determine whether air quality is bad, moderate, or good. Regarding the MSE and the attribution to these thresholds, the regression did not provide predictions convincingly for all three states. Better results are obtained with the more advanced LSTM model.

2 Introduction

Air pollution is one of the greatest environmental risks to human health (World Health Organization [WHO], 2021). Unclean air containing an abundant amount of ozone or particulate matter, for example, is a proven cause of illness and premature death (Bundesamt für Umwelt [BAFU], 2021a). In the study by Laeremans et al., (2018, S. 86), conducted in the cities of London, Barcelona, and Antwerp, it was statistically proven that physical activity in these cities increased the average heart rate and made breathing more difficult. The reason for these aggravations is the polluted air in the cities. In 2018, 2'300 people died prematurely in Switzerland due to air pollution. (BAFU, 2021a). Not only human health is endangered by air pollution, but also entire ecosystems, animals and plants. Air pollutants over-fertilize the fragile ecosystem and impact vegetation. Pollutants in the air are responsible for duck failures and contribute greatly to climate change (BAFU, 2021b).

The aim of this project is to analyze a historical dataset of air pollution from the USA. The entire analysis is done for Arizona, Kansas and Florida. The dataset included 127'276 datapoints for the daily CO (Carbon Monoxide), NO2 (Nitrogen Dioxide), O3 (Ozone) and SO2 (Sulfur Dioxide) information of each state. First, the data is analyzed with two models, one based on Recurrent Neural Networks (RNN) and the other with multiple linear regression. Second, a prediction about the data is made from 2015 onwards. Finally, a threshold regarding good, moderate and poor air quality is proposed for each state.

3 Theoretical Framework

This chapter treats the theoretical approach for the models which will be used in this paper to get an understanding of vocabulary used.

3.1 OLS Regression

Regression analysis is one of the most important areas in statistics and machine learning. Many regression methods, such as linear regression, are available. Regression looks for relationships between variables. Typically, regression is used to answer whether and how some phenomenon influences the other or how several variables are related. Therefore, a regression is also useful to forecast a response using a new set of predictors. One of the main advantages is the ease with which the results can be interpreted (Stojiljkovic, 2021). Linear regression is a statistical technique that is used to fit a straight line through the data (see Figure 1).

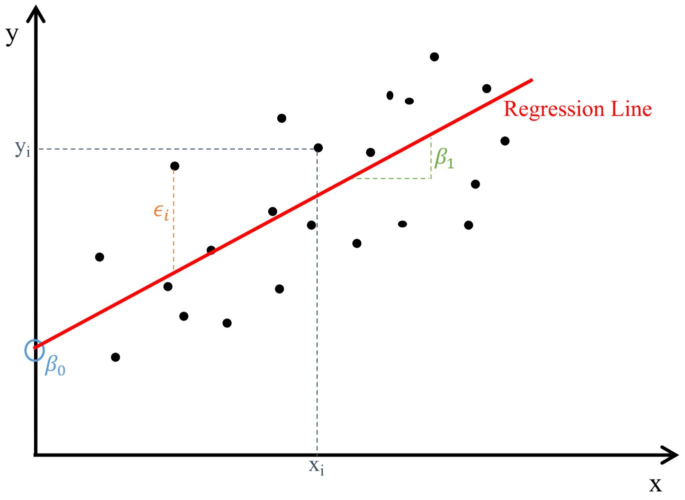

Figure 1: Simple linear regression (own representation based on Bachmann, 2020, p. 8 ff.)

This estimator produces a function that can be used to determine the value of the dependent variable for each value of an independent variable. The straight line is defined by the equation:

$$Y = \beta_0 + \sum_{i=1}^{n} (\beta_i X_i) + \varepsilon$$

With:

$Y = Dependent\ variable$

$\beta_0 = Y - intercept$

$\beta_i = Regression\ coefficients$

$X_i = Independent\ variable(s)$

$\varepsilon = Random\ error\ term$

Where Y is defined as the target size. Y can be figured out with the help of independent variables (X_i), the intercept (β_0) and the slope coefficients (β_i). The coefficients are the values by which the target size increases when the independent variables increase by one. The residuals (ε), or random error terms, are the unexplained portion of a dependent variable (StatisQuo, 2018). A regression is called multiple linear regression if the number of regression coefficients is larger than one.

The so-called Ordinary Least Squared (OLS) method is a specific technique which is based on the linear regression theory, which accounts for the sum of the squared deviations of each data point from the line. The OLS aims to minimize this sum to the smallest possible value to determine the best fitting line (Fox, 2015, p. 83 ff.).

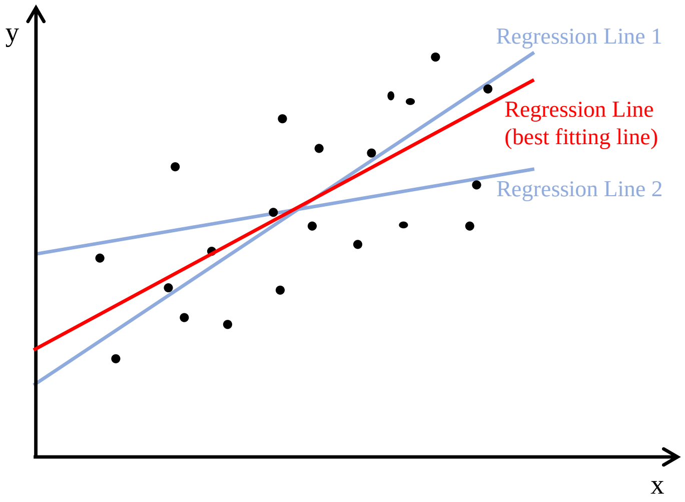

Figure 2: Simple linear Regression with OLS (own representation based on Bachmann, 2020, p. 16)

All three lines in Figure 2 were created by a linear regression and separate the data differently. But only OLS can determine the best fitting line. The main advantages of this method are, that it works well for linearly separable datasets and that overfitting can easily be handled. This is especially then the case when further dimensionally reduction techniques are used. Within the linearity lie not only the advantages but also the disadvantages. The main disadvantage is that this technique assumes linearity between the dependent and the explainable variables, which is not always the case in reality. Furthermore, outliers as well as multicollinearity have high influence on such calculations (Waseem, 2022). The formulas to apply the OLS for a simple regression equation such as $y_i = b_0 + b_1 x_i + e_i$ looks as follows (Newbold et al., 2020, p. 431 f.):

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

With:

$b_1 = Slope\ of\ the\ line$

$b_0 = Y - intercept$

$x_i, y_i = Variables$

$\bar{x}, \bar{y} = Mean\ of\ variables$

3.2 Recurrent Neural Networks

Traditional neural networks start their thinking from the beginning and have no persistence. This is the most significant weakness of traditional neural networks - they have no memory. Traditional neural networks are not able to complete the sentence "My name is ...", because they do not take into account the two previous words. However, as the name RNN suggests, they can complete the sentence and account for the words "My", "name" and "is" (and their context) for the prediction of the last word (a name). RNN are networks with loops that have a memory. The difference between RNN and traditional neural networks is that RNN are multiple copies of the same network, each passing a message to its successor (Olah, 2015).

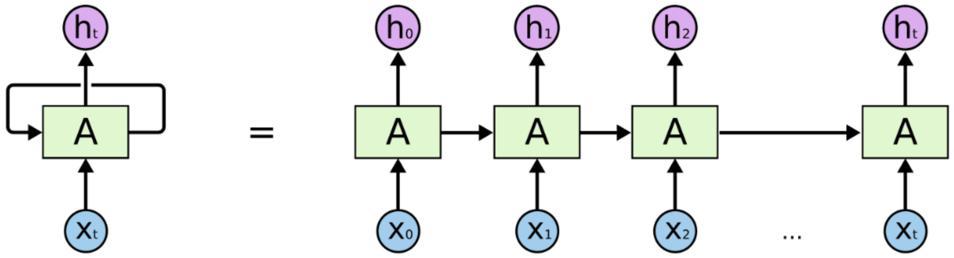

Figure 3: an unrolled recurrent neural network (based on Cola, 2015)

Figure 3 represents the structure of an RNN. The input is specified with the variable x_t, A (a part of the RNN) calculates the value h_t, which is integrated into the next part of the RNN by means of a loop. This chain can be written in a mathematical formula (Olah, 2015):

$$h\ (t) = f(h(t-1), x; 0)$$

This chain-like nature has achieved impressive successes in recent years including speech recognition, translation, captioning, etc. However, a large part of the success is due to Long-Short Term Memory (LSTM), a special type of RNN (Olah, 2015).

LSTM models are a refined variant that was developed specifically to handle the problem of long-term dependency. LSTM modelling also creates repeating modules of neural networks, but these have four layers that interact with each other in a special way (Mwangi, 2018).

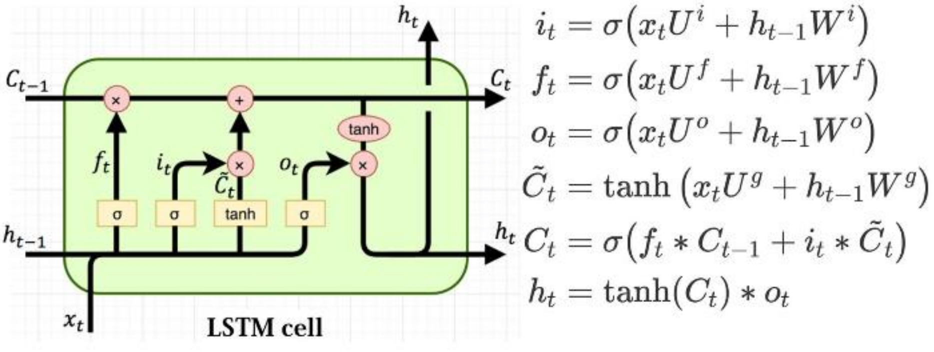
h_t
C_{t-1}
C_t
×
+
tanh
f_t
i_t
×
$\tilde{C}_t$
o_t
×
σ
σ
tanh
σ
h_{t-1}
h_t
x_t
LSTM cell
$i_t = \sigma(x_t U^i + h_{t-1} W^i)$
$f_t = \sigma(x_t U^f + h_{t-1} W^f)$
$o_t = \sigma(x_t U^o + h_{t-1} W^o)$
$\tilde{C}_t = \tanh(x_t U^g + h_{t-1} W^g)$
$C_t = \sigma(f_t * C_{t-1} + i_t * \tilde{C}_t)$
$h_t = \tanh(C_t) * o_t$

Figure 4: Structure of a LSTM cell and equations (Varsamopoulos et al., 2018, p. 4)

The cell state (C_t), the upper horizontal line in Figure 4, is the most important element, as information flows through it. By means of three gates (σ) it is possible to regulate this flow of information by adding or removing information and thus, protecting and controlling the cell state. This regulation is based on a sigmoid neural net layer, which has values between zero and one, determining how many components of information are allowed to pass through the gate. A value of one means that all information is allowed through, while a value of zero removes all information. Using the LSTM model, the flow of information goes through four steps. First, the forget gate layer decides which information from the previous cell is not relevant and can be removed from the cell state. In a further step, it is decided which new information will be stored in the cell state and consists of the decision from the input gate, which decides which values should be updated and enter the cell state. After that, a vector ($\tilde{C}_t$) is formed by means of $tanh()$, which prevents the distribution of the gradient from the problem of vanishing and addresses new candidate values. The combination of the input gate and the vector is then used as an update to the state. Finally, the decision is made about which output should be generated, which is based on the value from the cell state but is filtered again using the sigmoid gate to decide which information should be passed on to the next cell state (Olah, 2015).

3.3 Mean Squared Error

A wide variety of forecasting models can be developed for the most diverse areas of application with the help of the models described above. One method to measure the quality of forecasts is the so-called Mean Squared Error (MSE) technique, which is one of the best known and most used. The MSE calculates the squared error of the predicted value, which gives outliers a higher weighting in the assessment and thus, helps to filter them out (Padhma, 2021). Mathematically, the formula is as follows:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2$$

With:

$MSE = Mean\ squared\ error\ value$

$y_t = Actual\ value$

$\hat{y}_t = Forecasted\ value$

$n = Total\ number\ of\ observations$

The disadvantages of this measure are that the MSE is scale-dependent, which means that it is not meaningful for comparing different measures. Furthermore, it is not possible to distinguish whether one large error or many small errors are present, as the distinction is not made (Padhma, 2021).

4 Literature Review

In recent years, various investigations and studies have addressed the problems of air pollution and possible applications of deep learning tools for predictions. One such paper has been published by Athira et al. (2018, p. 1395), which also used RNN and LSTM as well as a so-called Gated Recurrent Unit (GRU) method to analyze the pollutant with the dataset obtained by AirNet. This dataset includes data on air quality with a sample size of over 10 million individual values. Athira et al. (2018) constructed and trained the models for 1000 epochs by shifting layers. Moreover, they conducted an analysis based on the MSE for a time period between April 2015 and September 2017. Within this study the authors confirmed that all three models can predict the future with a MSE value of 0.1661 for a LSTM model as the best and of 0.6569 for a GRU model as the worst (Athira et al., 2018, p. 1401 f.).

Also, Sagar et al. (2020, p. 241) analyzed air pollution and conducted a study for different cities of India. They obtained the hourly air pollution data from a wireless sensors network and from the Indian government, which provides real time data. For prediction, LSTM RNN was applied in this paper. During the period April 3 2019 to April 14 2019, gases such as CO, NH3, NO2, and SO2 were studied, to mention a few. The results conclude that good estimates for the future are feasible, but the sensor quality must be considered in such models. With their model, a prediction accuracy of ± 15-20% could be reached (Sagar et al., 2020, p. 246 f.)

5 Results

In this chapter the generated results are presented. The first subchapter contains the result for the OLS regression prediction, followed by the LSTM results in the second subchapter. The final subchapter demonstrates how the threshold for good, moderate, and bad air quality were defined and how well the models performed. The discussed results involve all three states whereas the presented pictures are limited to the state of Florida. In the appendix further visualizations for the other two states are presented.

5.1 OLS Regression

The results based on the MSE reveal that for Florida, the prediction made for the timeframe 2015 onwards, are the closest with a value of 2.8055 compared to the effective values. For Kansas a MSE value of 3.2156 could be achieved, whereas Arizona exhibits the worst result with a MSE value of 8.0872. Since the most accurate prediction was possible for the state of Florida applying the regression method, the actual and predicted datapoints are compared in Figure 5. The graphs of the other two states can be found in the appendix 8.2.

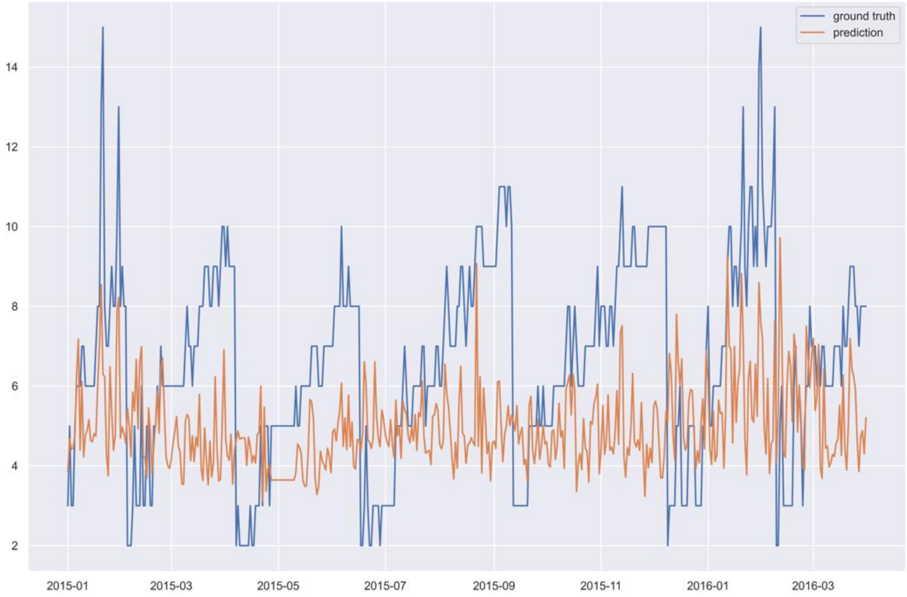

Figure 5: OLS for Florida (own representation)

5.2 LSTM

The results for the LSTM models contain two different approaches for each state. The first results were achieved with the following random parameters: two layers with 128 neurons each, five epochs and a batch size of one. The best result was detected for Kansas with a MSE value of 0.0131. The MSE value for Florida is 0.0765 and for Arizona 0.1176. A visual representation is displayed in appendix 8.3.

Further, the model for each state was tuned with three combined loops. The epochs (10, 25) were the first loop, the batch size (2, 4, 8) was the second loop and the neurons (32, 64, 128) were the third loop. In total 18 different parameter combinations were examined for each state. The best and hence selected parameters for each state are as follows:

	Layers	Neurons	Number of Epochs	Batch Size
Arizona	2	128	25	2
Florida	2	128	25	4
Kansas	2	128	25	2

Table 1: Tuned parameters (own representation)

For all three states the results based on the MSE values could be improved. The best result was still achieved for Kansas with a new MSE value of 0.0007. For Florida a MSE value of 0.0239 was reached and for Arizona, a value of 0.0014. An overview of all MSE is displayed in Table 2.

	Random	Tuned
Arizona	0.1176	0.0014
Florida	0.0765	0.0239
Kansas	0.0131	0.0007

Table 2: MSE values for random and tuned LSTM models (own representation)

Figure 6 contains the actual and predicted values for Florida. The others can again be found in the appendix 8.4.

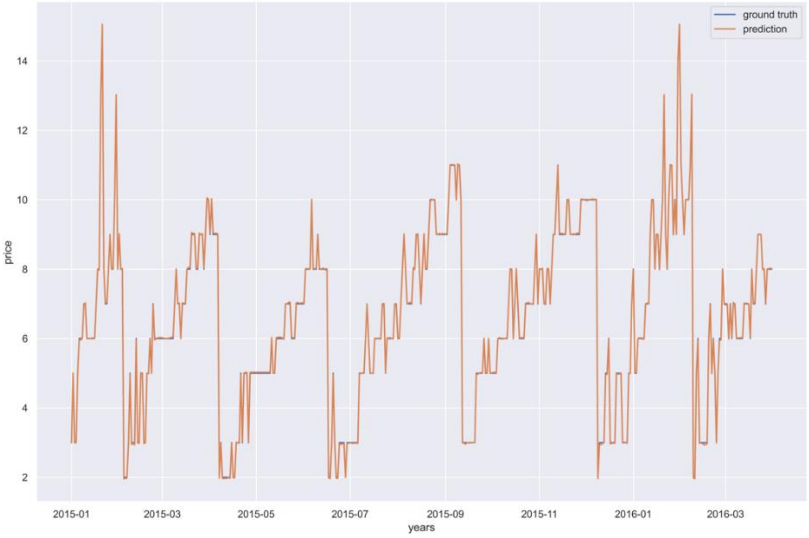

Figure 6: LSTM with tuned parameters for Florida (own representation)

5.3 Threshold

Air quality was divided into three categories: good, moderate and bad. For good, the 25th percentiles and for bad, the 75th percentiles were defined. This was specified for each state. Based on the effectively measured values, the predicted values were subdivided to measure the quality of the predictions taken into account the three categories. Comparing actual and predicted values, with an average difference of -0.67%, the tuned LSTM model achieved the best results for Florida. In the appendix 8.5 a visual representation of the actual and predicted air quality attribution is presented for each state.

6 Conclusion

As expected, the LSTM model had a better accuracy than the regression model. It was revealed that simple approaches such as OLS are less suitable for complex predictions topics like air quality. Instead, it is necessary to use more advanced models to obtain convincing results.

If the results are further compared to those of Athira et al. (2018), it can be stated that with the model used in this project, a higher accuracy could be achieved across all states. However, the amount of data considered plays a crucial role. Ahtira et al. (2018) considered a higher amount of data in their analysis whereas Sagar et al. (2020), on the other hand, used less data than in their paper and were able to significantly increase the accuracy of their predictions.

Future research should include determining the best parameters for the models. Furthermore, it would be useful if a uniform definition of the threshold is developed in order to ensure the comparability of the results.

7 Bibliography

Athira, V., Geetha, P., Vinayakumar, R., & Soman, K., P. (2018). DeepAirNet: Applying Recurrent Networks for Air Quality Prediction. *Procedia Computer Science, 132*. 1394–1403. https://doi.org/10.1016/j.procs.2018.05.068

Bachmann, O. (2020). *Econometrics - OLS Estimation of Classical Linear Regression Model (Newbold 11.1-11.4, 12.1-12.3) [lecture script]*. Zurich University of Applied Science, Department Quantitative Finance.

Bundesamt für Umwelt [BAFU] (2021a). *Auswirkungen der Luftverschmutzung auf die Gesundheit*. https://www.bafu.admin.ch/bafu/de/home/themen/luft/fachinformationen/auswirkungen-der-luftverschmutzung/auswirkungen-der-luftverschmutzung-auf-die-gesundheit.html

Bundesamt für Umwelt [BAFU] (2021b). *Auswirkungen der Luftverschmutzung*. https://www.bafu.admin.ch/bafu/de/home/themen/luft/fachinformationen/auswirkungen-der-luftverschmutzung.html#:~:text=belasten%20die%20Volkswirtschaft.-,auf%20die%20%C3%96kosysteme,Vegetation%20und%20f%C3%BChrt%20zu%20Ernteausf%C3%A4llen

Fox, J. (2015). *Applied regression analysis and generalized linear models* (3. edition). Sage Pubn.

Laeremans, M., Dons, E., Avila-Palencia, I., Carrasco-Turigas, G., Orjuela, J. P., Anaya, E., Cole-Hunter, T., de Nazelle, A., Nieuwenhuijsen, M., Standaert, A., Van Poppel, M., De Boever, P., & Int Panis, L. (2018). Short-term effects of physical activity, air pollution and their interaction on the cardiovascular and respiratory system. *Environment International, 117,* 82–90. https://doi.org/10.1016/j.envint.2018.04.040

Mwangi, B. (2018). *An Intro Tutorial for Implementing Long Short-Term Memory Networks (LSTM).* https://heartbeat.comet.ml/a-beginners-guide-to-implementing-long-short-term-memory-networks-lstm-eb7a2ff09a27

Newbold, P., Carlson, W. L., & Thorne, B. M. (2020). *Statistics for Business and Economics* (9. edition). Pearson Education Limited.

Olah, C. (2015). *Understanding LSTM Networks*. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Padhma, M. (2021). *End-to-End Introduction to Evaluating Regression Models.* https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/

Sagar, V., B., Sreenidhi, R., Ranjani, R., & Rajasekar M. (2020). Air Quality Forecasting using LSTM RNN and Wireless Sensor Networks. *Procedia Computer Science, 170.* 241–248. https://doi.org/10.1016/j.procs.2020.03.036

StatisQuo (2018). *Lineare Regression und Anwendung in Python*. https://statisquo.de/2018/03/23/lineare-regression-und-implementierung-in-python/

Stojiljkovic, M. (2021). *Linear Regression in Python*. https://realpython.com/linearregression-in-python/

Varsamopoulos, S., Bertels, K., & Almudever, C., G. (2018). *Designing neural network based decoders for surface codes.* https://www.researchgate.net/publication/329362532_Designing_neural_network_based_decoders_for_surface_codes

Waseem, M. (2022). *How To Implement Linear Regression for Machine Learning?.* https://www.edureka.co/blog/linear-regression-for-machine-learning/

World Health Organization [WHO]. (2021). *Ambiente (outdoor) air pollution.* https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

8 Appendix

8.1 Github

All our data and Jupyter Notebooks can be found on Github:

https://github.com/bittelandreas/USAirPollution

8.2 OLS Results

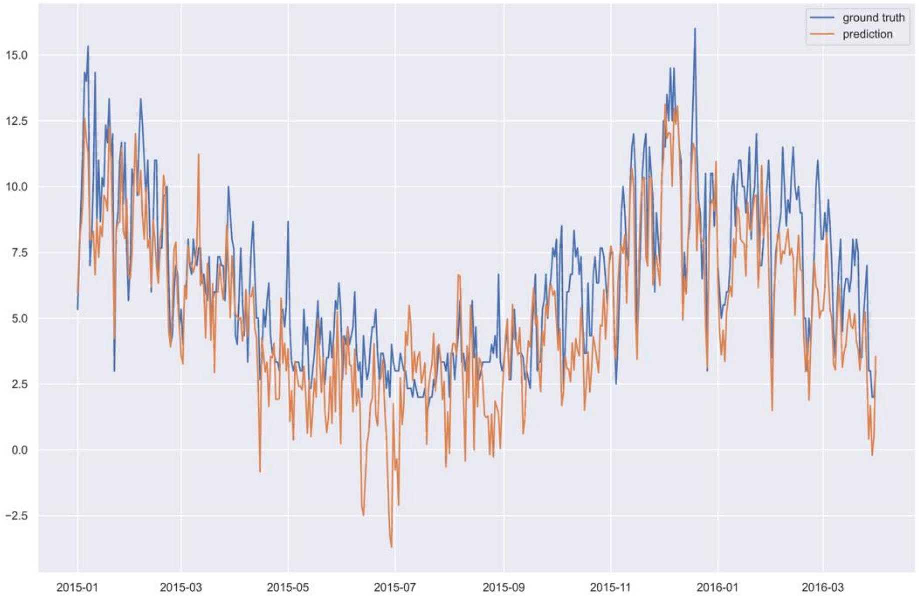

Figure 7: OLS for Arizona (own representation)

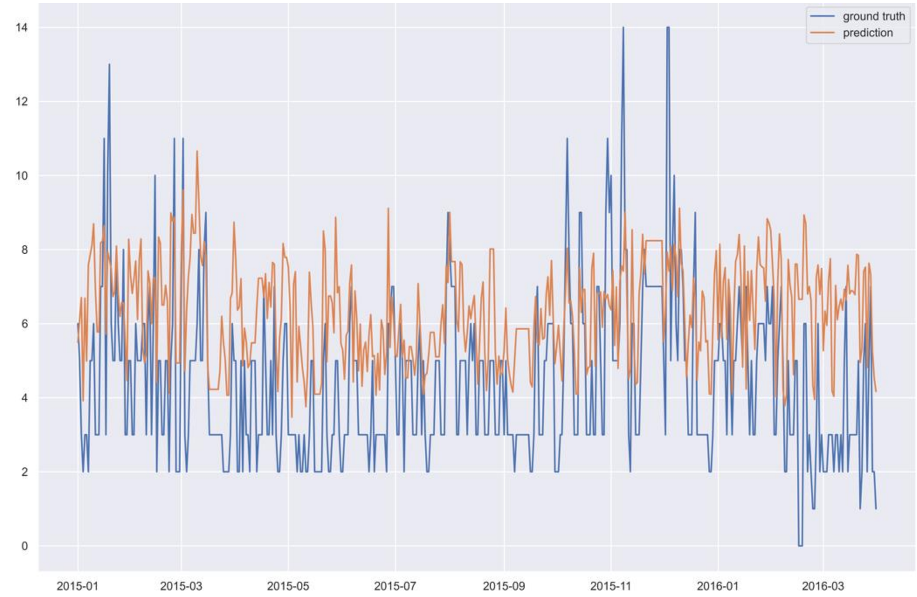

Figure 8: OLS for Kansas (own representation

8.3 LSTM results with random parameters

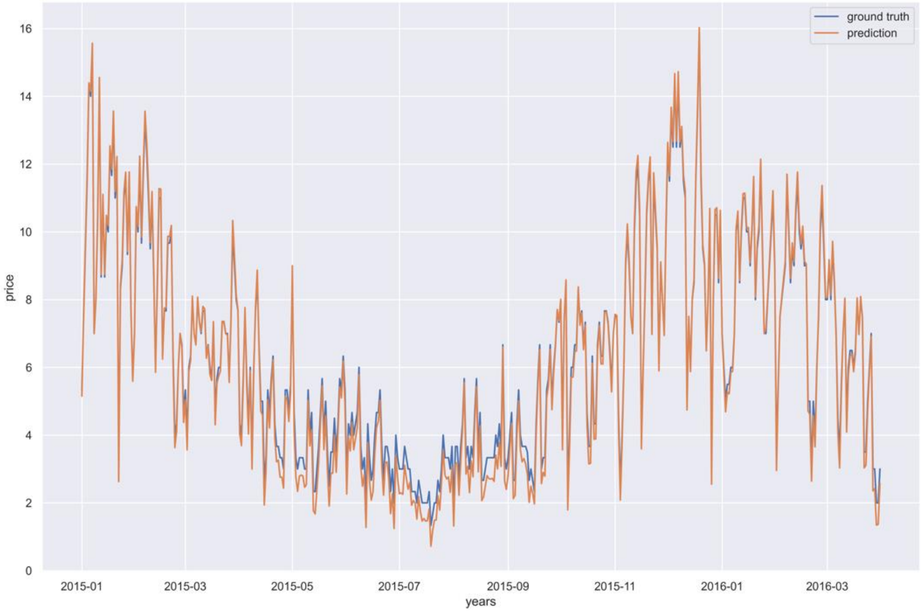

Figure 9:LSTM with random parameters for Arizona (own representation)

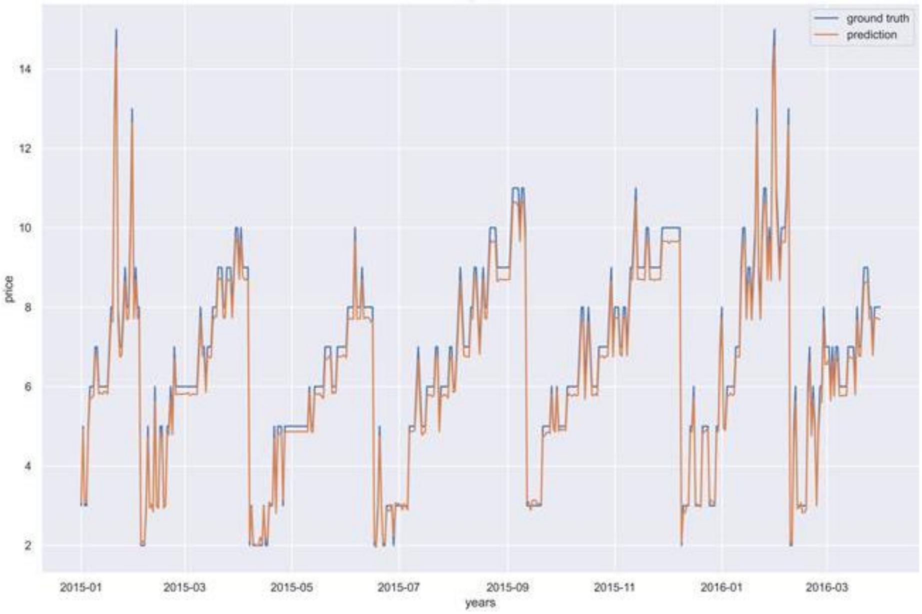

Figure 10: LSTM with random parameters for Florida (own representation)

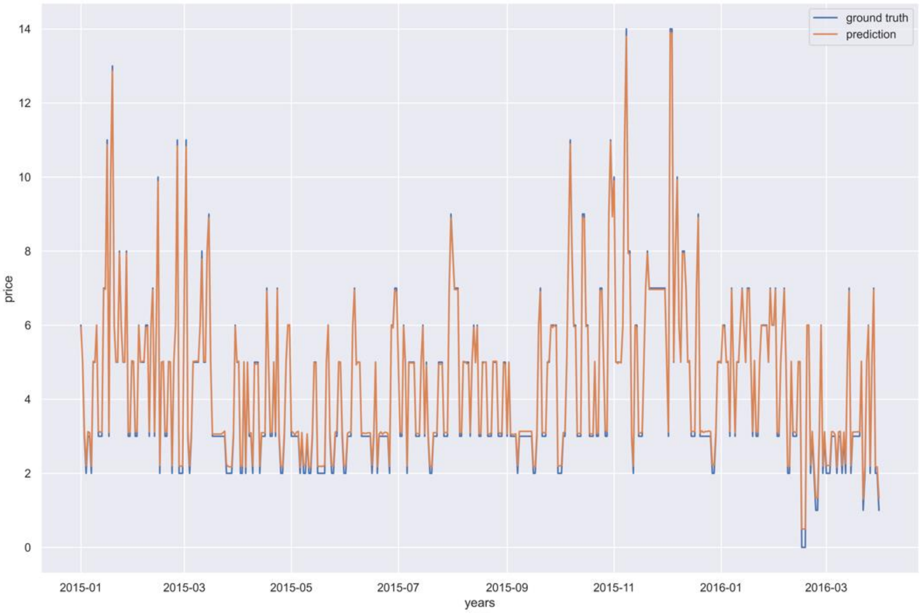

Figure 11: LSTM with random parameters for Kansas (own representation)

8.4 LSTM results with tuned parameters

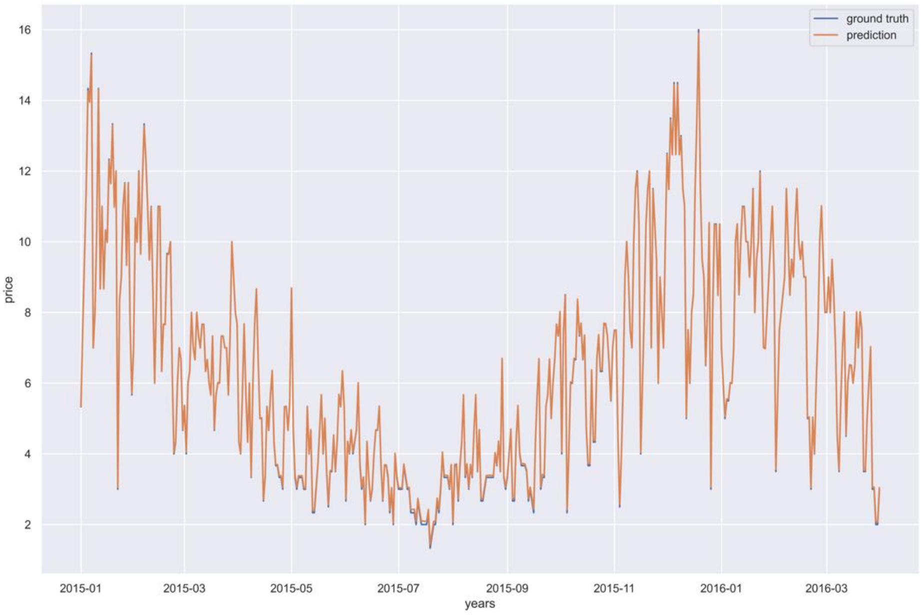

Figure 12: LSTM with tuned parameters for Arizona (own representation)

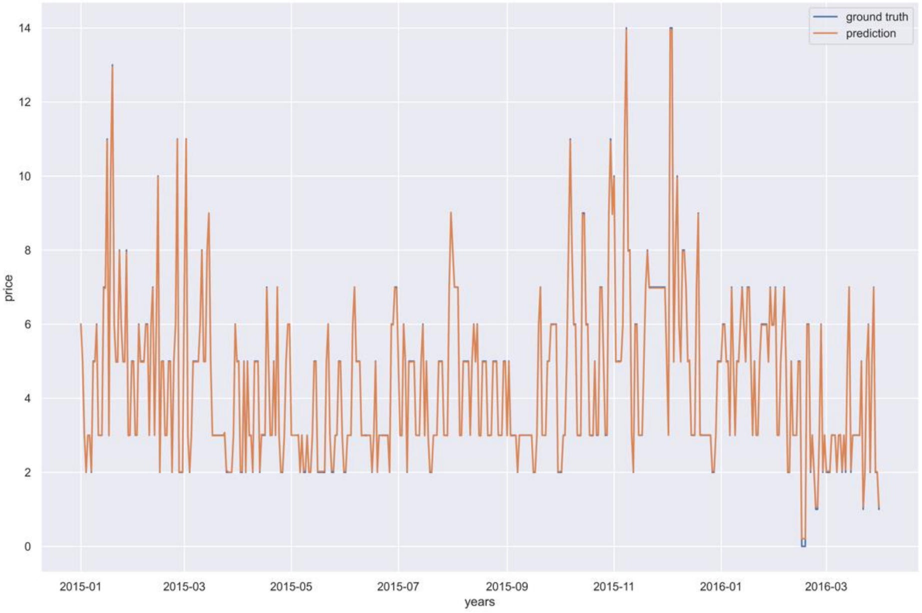

Figure 13: LSTM with tuned parameters for Kansas (own representation)

8.5 Threshold results

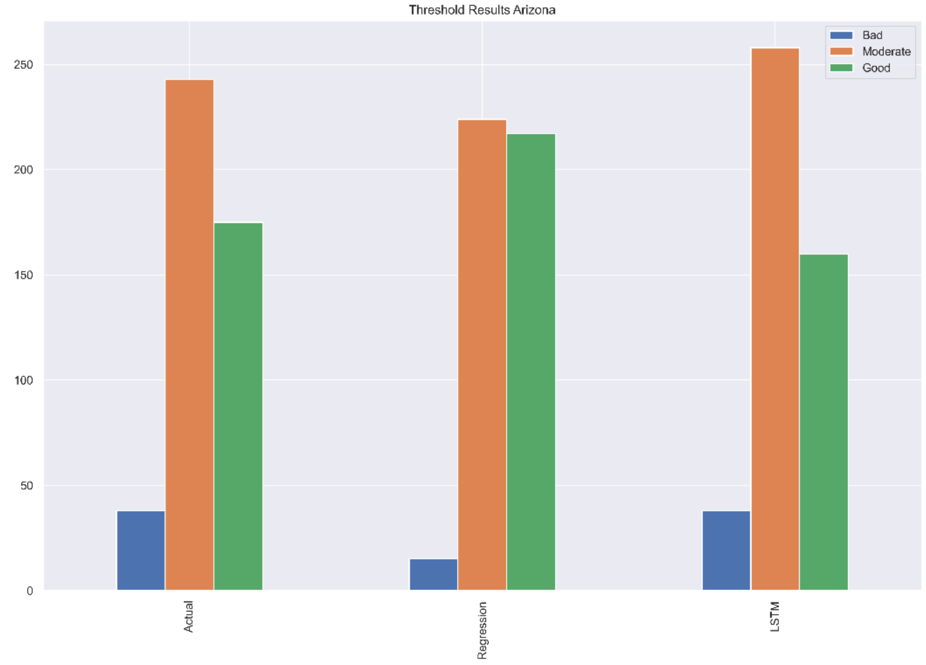

Figure 14: Air quality results for Arizona (own representation)

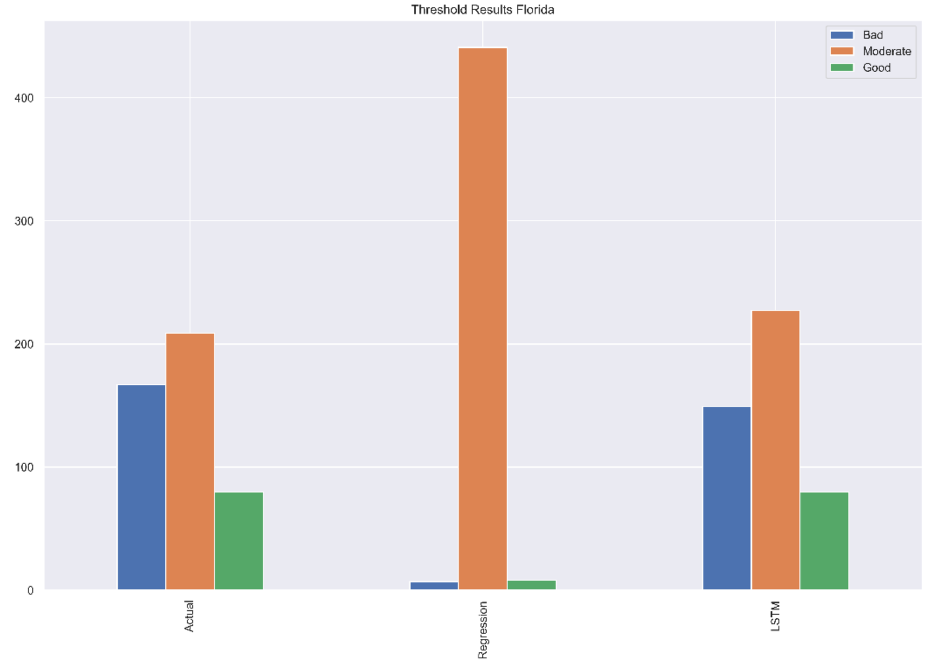

Figure 15: Air quality results for Florida (own representation)

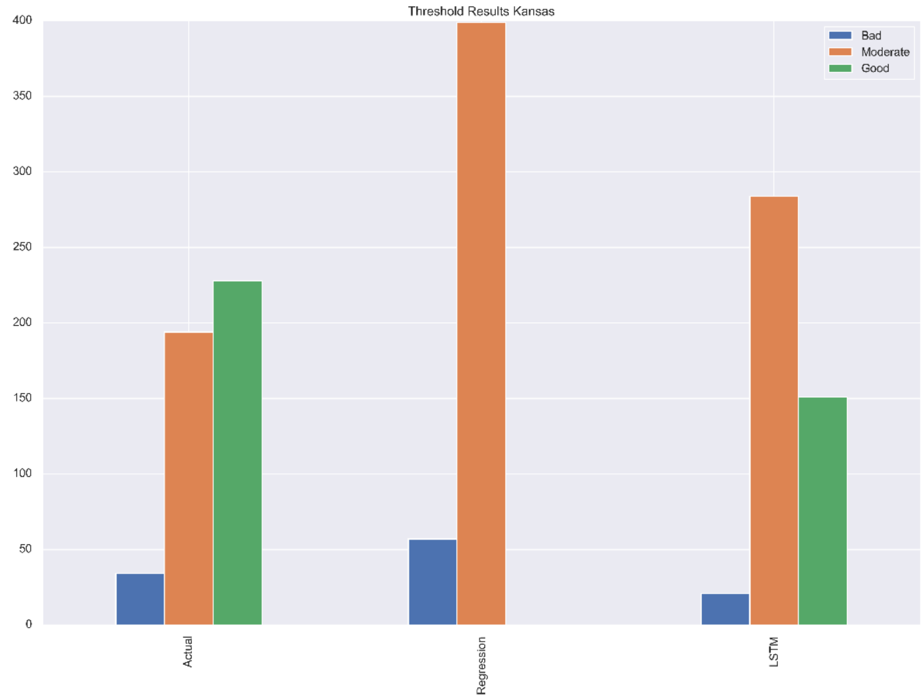

Figure 16: Air quality results for Kansas (own representation)