Predicting MP 2.5 levels in Temuco, Chile using Deep Learning

Tomás Rojas
Matías Montagna
Alonso Utreras
Dept. de Ingeniería Matemática
Universidad de Chile
Santiago, Chile

***Abstract*—Temuco is one of the most highly wood-smoke-polluted cities in the world [1]. The $PM_{2.5}$ levels registered are high due to the high usage of chimneys. It has become important to predict pollution levels in order to prevent activities in schools and other institutions since this may cause health problems. This work presents a forecasting model to predict $PM_{2.5}$ levels for the mentioned city. Deep Learning models were used, showing how do different architectures perform and how important are different factors like temperature, relative humidity, atmospheric pressure, wind direction and wind speed.**

***Index Terms*—**

Atmospheric pollution is currently one of the most important environmental problems at global scale [2]. Some estimations point that 6.5 millions of deaths are associated with indoor and outdoor air pollution. Approximately 92% of the global population breaths toxic air [3]. Temuco is no exception to this, where pollution is generated by burning wet wood, which is a cheap way of reaching a comfortable level of temperature and therefore is being used by the poor inhabitants of the region [4].

Watching $PM_{2.5}$ historical levels in the air, it has been observed that this phenomena occurs mostly during winter [5]. There has been some governmental programs with the objective of reduce firewood dependency [6]. In this context, predicting future air pollution levels may be useful for a right application of these laws. Deep learning can be a tool for helping accomplish this.

I. Related work

A bibliographic review on machine learning to predict pollution has been done by [7]. They showed that different approaches have been tried in order to find the most suitable for the task.

Also sequential modules have been tried, like the one described in [8] where a model with two fully connected layers and a LogSigmoid function is used to predict average PM_{10} in a day, given temperature, relative humidity, wind direction and wind speed. This approach was used for every station in the city.

Another work by [9] used Bi-DLSTM to predict $PM_{2,5}$ levels using latitude, longitude and time. This model was able to predict pollution levels for regions between the stations where data was taken from. It used data from the past and from the future.

[10] made estimations of $PM_{2.5}$ of yearly or seasonal concentrations using Spatial Back Bropagation Neural Networks (SBPNN), using as input Wind speed, Relative humidity, Temperature, Air Pressure, Precipitations, Sunshine Hours, Aerosol Optical Depth, Land Use, Point Emission Use, among others. This study had one of the best results, but it must be considered that plenty of data were used in order to accomplish this results.

We chose to experiment with different models and varying inputs. The data used come from three [5] stations. These include: relative humidity, temperature, atmospheric pressure, wind speed and direction, date, hour and $PM_{2.5}$.

II. Methodology

We downloaded Time Series data in .csv format from the [5]. The csv files are as follows:

DATE;HOUR;VALUE;
040331;0200;;

This format was analogous for each measurement (i.e. temperature, pressure, pollution), so we had to place all data together in order to feed a NN. A lot of data was missing, so we had to pre-process them using different techniques, like applying filters like in figure 1 and pooling it with mean functions. The values of the nine stations were pooled, so missing data was not a problem.

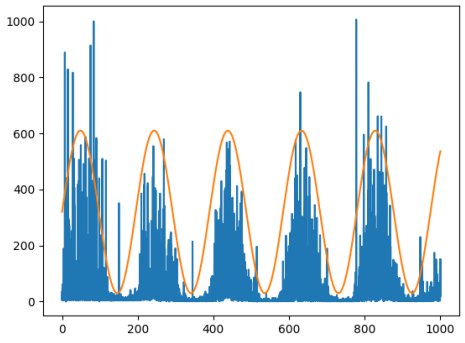

Figure 1. Filtering $PM_{2.5}$ anomalies in the data

The $PM_{2.5}$ empiric pdf is shown in the figure 2, showing that the higher the PM value, the less frequent this is. After testing this demonstrated to be an important bias for the results of our models.

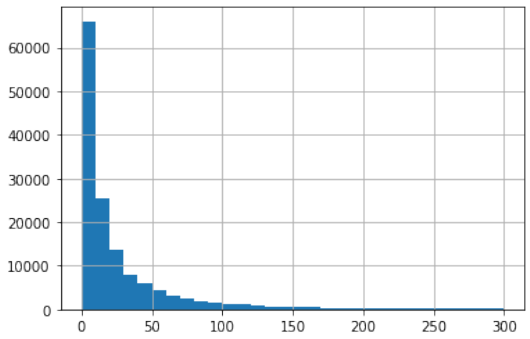

Figure 2. Histogram of $PM_{2.5}$ levels

After the data was merged, we experimented with different models and architectures. The best performance was acquired by a Sequential model using three fully connected layers of dimensions (6 x 128 x 128 x 1), whereas the first activation is a ReLu function and the second one a LogSigmoid. The same as shown in the literature.

Regarding the hyper parameters, we tried with different batch sizes, but the differences were not significant. In order to work by days and considering the data was by hours, we used batches of size 24. The optimizer used was Adam, the loss function MSE. SGD and L1 did not perform better as the former ones. We experimented with the number of epochs. A dropout rate of 0.25 was used in order to prevent overfitting.

III. Results and Discussion

The model used was a Triple Model consisting on three fully connected layers. In order to contrast results and find relations on the features used in the model, several iterations were made were the model was fed all but one or more features. The results are shown in table I. The evolution of the loss function during training is represented in figure 5.

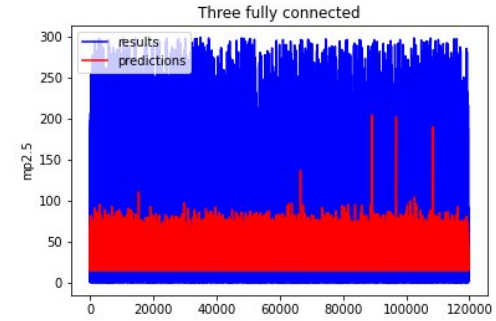

Figure 3. Model with all features and filtered data.

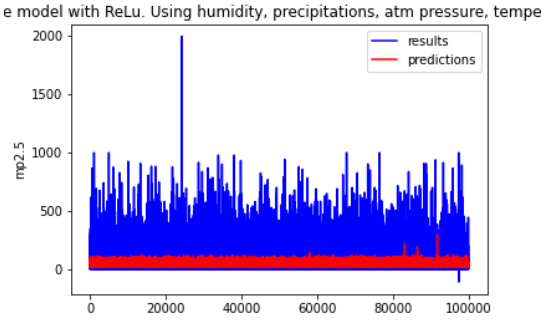

Figure 4. Model without both wind metrics

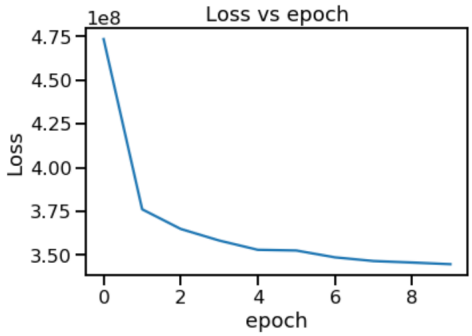

Figure 5. Triple Model's Loss v/s epoch

	All metrics	Without temperature	Without precipitations	Without humidity	Without temperature	Without pressure	Without both wind features
R2 score	0.21084	0.11969	0.19629	0.18133	0.11969	0.17813	0.13504

Table I
R2 scores with different features

It was seen that every feature had non-negligible relevance, having impact on the R2 score given its presence/absence. Having said that, there were three features that stood out among the others based on its impact in performance of the model: temperature, and the two features related to the wind (wind direction, wind velocity). This can be explained by the correlation between winter and the contamination in Temuco. When it gets colder, fireplace activity increases significantly. Also, when wind velocity is low, the pollutants in the air do not

disperse and that leads to more pollution. But also it could have been that the correlation between those two variables was high so it didn't matter that one was not fed to the model because another one cointained the mayority of its info (Figure 6).

Several activation functions were used, but the ones that gave that were the most consistent on giving good results were the ReLu function and the Log Sigmoid function.

IV. Conclusions

Even though we used different number of epochs, there were not significant differences after the eighth epoch, even in the validation sets.

Given the nature of this problem, it is difficult to predict. Our model failed when $MP_{2.5}$ levels were very high for short periods of time, this is reasonable because the ability to predict them was not only dependent in the statistics behind this events, but also dependent on the smoothness of the functions involved in the model. Given the nature of these data, the higher the value the less likely it was. Even though, predictions were good for low $MP_{2.5}$ levels days.

We can conclude that data pre-processing might be the most decisive part of the workflow. Our results are relatively flat because most of the values are close to zero as shown in figure 2. All features are significant to improve prediction values. Our datasets had mistakes (such as negative values, or impossible high measurements). Even with filtering, data were still not predictable for a Neural Network. Given this, it is really important for governmental institutions to acquire accurate data.

The LSTM did not show better results. It produced more spikes with its predictions, but results were ill defined.

According to our experiments and the literature, both the ReLu and LogSigmoid activation functions work better for these cases.

V. Further Improvements

For further improvements, another loss function should be used, strongly penalizing when predicted data is much less than real data, so the NN could try to replicate eventual peaks in $PM_{2.5}$ levels.

Also, other features could be added: localization, datetime, vegetation level, government policies, wood price, poverty levels and house isolation levels.

The code we used to mine/merge the data, train and evaluate our models is available at https://github.com/AlasAltum/Proyecto_ML_MP2.5

Bibliography

[1] P. A. Sanhueza, M. A. Torreblanca, L. A. Diaz-Robles, L. N. Schiappacasse, M. P. Silva, and T. D. Astete, "Particulate air pollution and health effects for cardiovascular and respiratory causes in temuco, chile: a wood-smoke-polluted urban area," *Journal of the Air & Waste Management Association*, vol. 59, no. 12, pp. 1481–1488, 2009.

[2] B. Bishoi, A. Prakash, V. Jain, *et al.*, "A comparative study of air quality index based on factor analysis and us-epa methods for an urban environment," *Aerosol and Air Quality Research*, vol. 9, no. 1, pp. 1–17, 2009.

[3] V. Luchkevich, M. Galina, and V. Filatov, "Health-related quality of life assessment when implementing ecology programs," in *E3S Web of Conferences*, vol. 176, p. 04012, EDP Sciences, 2020.

[4] "The world's worst air is in this south american city thanks to poverty," Jul 2020.

[5] SINCA, 2020.

[6] A. Cortes-Fuentes and B. Rismanchi, "Residential energy efficiency in chile: Policies to reduce firewood dependency/eficiencia energética residencial en chile: políticas para reducir la dependencia de la leña," *Revista ESTOA*, vol. 9, no. 17, pp. 57–69, 2020.

[7] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *International Journal of Environmental Science and Development*, vol. 9, no. 1, pp. 8–16, 2018.

[8] M. G. Cortina-Januchs, J. Quintanilla-Dominguez, A. Vega-Corona, and D. Andina, "Development of a model for forecasting of pm10 concentrations in salamanca, mexico," *Atmospheric Pollution Research*, vol. 6, no. 4, pp. 626–634, 2015.

[9] W. Tong, L. Li, X. Zhou, A. Hamilton, and K. Zhang, "Deep learning pm 2.5 concentrations with bidirectional lstm rnn," *Air Quality, Atmosphere & Health*, vol. 12, no. 4, pp. 411–423, 2019.

[10] W. Wang, S. Zhao, L. Jiao, M. Taylor, B. Zhang, G. Xu, and H. Hou, "Estimation of pm2. 5 concentrations in china using a spatial back propagation neural network," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

VI. FIGURES

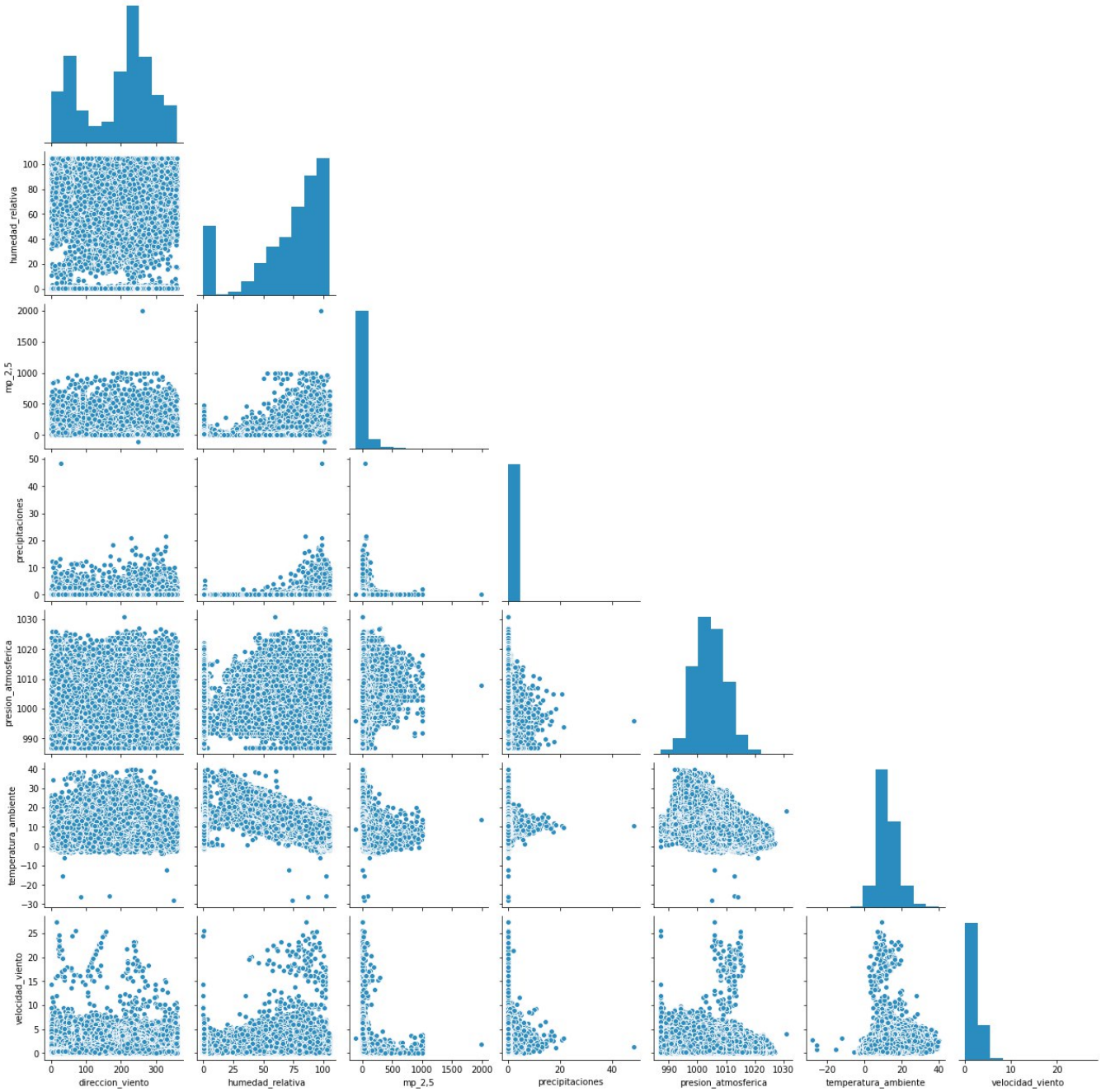

Figure 6. Correlation of all features