Data Cleaning and Feature Engineering

J. Sulo
Institute for Atmospheric and Earth System Research (INAR), University of Helsinki
juha.sulo@helsinki.fi

1 INTRODUCTION

One of the most important steps for any scientific study are data cleaning and feature engineering before the initial analysis. Data cleaning is the step in which the data one intends to analyze is first cleaned of data that may make the analysis more difficult or even impossible. This includes several different things such as removing duplicate observations and unwanted outliers or fixing structural errors such as poorly named categories. Of course, one of the key elements of data cleaning is also dealing with any missing data points. Data cleaning is a key element in all data analysis and a most necessary one for accurate and meaningful results.

In addition to this, feature engineering, or the process in which new, more meaningful features for the analysis are created, is important. This can be achieved by creating new interaction features, combining similar but sparse classes, removing unused variables or creating dummy variables. According to a recent survey, data scientists spend up to 70% of their time on these two steps [6].

In this study, I will review some of the more common data cleaning and feature engineering techniques in atmospheric sciences and provide examples from Particle Size Magnifier (PSM) data [4] and other miscellaneous data science sources. The use of proper data cleaning and feature techniques has a marked effect on the results of the analysis and the results for data with and without cleaning is compared. The use of domain knowledge is shown to be useful as interaction features are introduced that estimate particle growth rates in the atmosphere. Lastly, the importance of data cleaning and feature engineering is discussed and the role of machine learning in it.

2 DATA CLEANING IN ATMOSPHERIC SCIENCES

Data cleaning is perhaps the most important process in data analysis. Without proper cleaning of data, no matter how good the model or the analysis, the results won't be reliable. Data cleaning is necessary whenever there is raw data that needs to be analyzed. Well cleaned data should be have little noise, unwanted data points such as outliers should be removed and gaps in the data set should be filled as needed.

In atmospheric sciences, data cleaning is necessary when using measurement data either for analysis or as model input. In both cases the quality of the data is paramount to a good result. Measured data, be it from field or lab measurements, is prone to errors such as machine malfunctions, local emissions, power outages, inefficient storage, poor assumptions in signal management and inversions, and many more. All of these have to be accounted for in some way or another. Either by making assumptions that they are negligible or corrected for before the actual analysis.

In atmospheric sciences, most measurements come from field measurements, which are particularly susceptible to fluctuations due to changing weather conditions, power outages and others [3]. If the measurements are long term measurements, it adds the added difficulty of inevitable device maintenance and replacement devices. This introduces perhaps the most common problem of atmospheric data sets, missing values. Dealing with missing data points is an important decision that can be done in many ways. A common solution is to leave these points out of the analysis, but this introduces problems with several statistical tests that require a continuous time series. Many statistical tests such as Sen's slope require an equally spaced and continuous data set, but if for some reason it is not applicable to average the data further, you must decide on how to fill in the missing data points. Of course, for some analyses simply ignoring missing data points can be enough, but particularly for trend analysis, continuous data sets are needed.

Gap filling method	Median	Mean	Standard deviation
Original	185.9	245.8	256.3
Zero-filled	20.4	150.7	233.7
Nearest non-missing	182.2	243.4	257.1
Linear regression	184.5	244.8	255.5

Table 1: Comparison of the different gap filling methods and the original dataset. The original dataset contains 52 704 data points, with 20 399, or 38.7% of data points missing. The data is one year (2016) of particle concentration data from SMEAR II station measured with the PSM.

There are several options for filling in the data points. A common, although flawed, technique is to just fill in the missing data points with zeroes. This can however, skew statistical tests even further and depending on the amount of missing data points, make calculating means and medians meaningless. A better solution is to use some sort of interpolant to fill in the values. The simplest choice is to simply fill in the missing values with the previous or next data point. A linear regression model or a spline interpolant gives a much better estimate of the missing values and skews statistical tests less. An example of the effect of different methods of the series mean, median and standard deviation is presented in Table 1. In general, it is good to keep an eye on statistical variables such as median, mean and standard deviation while performing data cleaning. Particularly in trend analysis, you don't want your statistical indicators to rapidly differ between the cleaned and uncleaned dataset. Some variance between the cleaned and uncleaned dataset is to be expected, especially if your signal-to-noise ratio is low.

In addition to gap filling, other important data cleaning steps include removing unwanted outliers and structural observations such as detecting break points in the data and combining different data categories. Typical unwanted outliers in field measurements are negative values and extreme values that are not possible based on domain knowledge. This also ties in with break points, where a clear and noticeable shift occurs in the data and a portion of the data has to be discarded because the values are several orders of magnitude different from the rest of the data set with no clear explanation to the sudden shift. It is important to take care in discarding observations, because outliers and extreme values can always hide new science and new results that may be very important to the analysis. A sudden spike in aerosol concentrations can be attributed to device malfunction, but it could also be a local pollution episode, and if your goal is to investigate local pollution, you should not discard those data points. However, if you wish to investigate long term trends and background activity, it may be better to discard the episodes which show a local pollution episode. Other important procedures are making sure the structure of the data is sound, with appropriate column names, and making sure there are no duplicate observations.

Finally, while most data cleaning targets can be considered noise, the traditional meaning of noise in measurements is perhaps the random electric noise in the signal. Signal noise is typically caused by the measurement device itself and does not represent any real phenomena. It is also often rapid in frequency and depending on its amplitude, can hide important yet subtle features in the data. Particularly if the signal-to-noise ratio is much larger than 1, noise can be smoothed with many smoothing methods. Common smoothing methods include calculating moving means or medians in the data, using a Gaussian-weighted moving average filter or exponential smoothing. Another popular method in atmospheric sciences is the local regression method, known as LOWESS (locally weighted scatterplot smoothing). Local regression fits simple models to local susets of the data to create a function that best estimates the deterministic portion of the data. This has the advantage of not requiring the analyst to specify a specific global model to use for the data, but it does make the technique somewhat more computationally expensive.

3 FEATURE ENGINEERING IN ATMOSPHERIC DATA ANALYSIS

Feature engineering is a common tool used in data science to increase model output, performance and strengthen analysis. Feature engineering commonly uses domain knowledge to create new interaction variables by combining existing variables, combining data sparse data classes or creating completely new variables from complementary data to help the analysis. It's also possible to remove possible trends in an attempt to separate two variable dependent on one another by removing the associated trend. Similarly, dimensionality reductions methods such as principla component analysis can be used to deal with correlated variables and provide new features for analysis.

Feature engineering in atmospheric data analysis is a stable tool that is used to powerful effect in analyzing different correlations and causes for atmospheric phenomena. Typical features engineered can be combinations of measurements simulating particle growth rates, sulphuric acid concentrations, or modelling complex atmospheric phenomena such as global temperatures. In atmospheric size distributions, adjacent size classes can be combined to increase data coverage and simplify the analysis. Domain knowledge can be used to create new variables, for example indicator variables to isolate only clear sky days in new particle formation studies [1]. Additionally, dummy variables can be created to improve analyses, for example new particle formation event classification [2] variable to study differences in new particle formation event and non-event days.

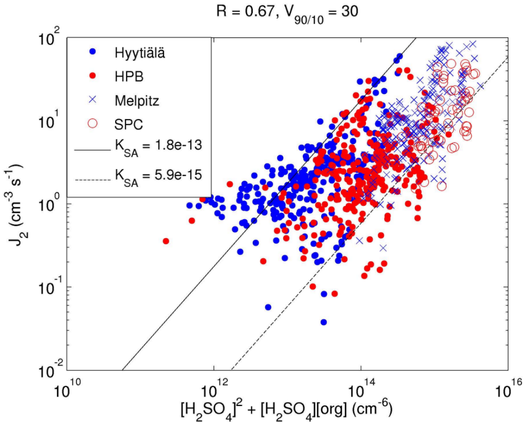

Figure 1: The neutral particle formation rate J_2 as a function of the new constructed feature, the sum of the sulphuric acid concentration and the product of sulphuric acid and organic vapour concentrations. The lines present the 90th and 10th percentile values of the coefficient $K_{\mathbf{SA}}$ (Paasonen, 2010)

An example of feature engineering is presented in [5], where measured concentrations of atmospheric vapors are used to estimate particle growth rates. The constructed interaction variable assumes that the gases used in the construction of the formula take part in the growth process in a steady state process. Figure 1 shows the constructed variable agrees quite nicely with the growth rate calculated directly from measured particle distributions. More complicated variables can be constructed as well, and this is particularly true for global atmospheric models.

4 THE ROLE OF MACHINE LEARNING IN DATA CLEANING AND FEATURE ENGINEERING

Data cleaning and feature engineering provide a significant challenge to deep learning methods because no data is ever the same and each dataset provides its unique data cleaning and feature engineering challenges. This requirement of domain knowledge makes it difficult to provide a general training set for data even within the same field. Data cleaning methods can also be developed for a specific dataset in mind, but validating these results is difficult. The black box nature of deep learning algorithms doesn't let us know the reasons why the algorithm removes certain data points or creates some new features for the analysis. This makes it very challenging to justify removing some features or data points, particularly because training these algorithms for data cleaning can be a very daunting task.

5 CONCLUSION

Data cleaning and feature engineering are the most time-consuming portions of most data analysis projects today. Deep learning algorithms or models, and data science tools in general could not be developed without proper data cleaning and feature engineering first. Data cleaning has grown into more prominence with the advent of big data and larger datasets becoming available in many fields. In atmospheric sciences, the availability of longer time series requires the development of more rigorous data cleaning methods to ensure the quality of the data used in analysis and the development of models. Similarly, feature engineering is required to provide more insights into ever more complex models and theories.

REFERENCES

[1] Dada, L., Paasonen, P., Nieminen, T., Buenrostro Mazon, S., Kontkanen, J., Peräkylä, O., Lehtipalo, K., Hussein, T., Petäjä, T., Kerminen, V.-M., Bäck, J., and Kulmala, M. Long-term analysis of clear-sky new particle formation events and nonevents in hyytiälä. *Atmospheric Chemistry and Physics 17*, 10 (2017), 6227–6241.

[2] Dal Maso, et al. Formation and growth of fresh atmospheric aerosols: eight years of aerosol size distribution data from SMEAR II, Hyytiala, Finland. *BOREAL ENVIRONMENT RESEARCH 10*, 5 (OCT 24 2005), 323–336.

[3] Kerminen, V.-M., Chen, X., Vakkari, V., PetÃďjÃď, T., Kulmala, M., and Bianchi, F. Atmospheric new particle formation and growth: review of field observations. *Environmental Research Letters 13*, 10 (sep 2018), 103003.

[4] Lehtipalo, K., et al. Methods for determining particle size distribution and growth rates between 1 and 3 nm using the Particle Size Magnifier. *Boreal Env. Res. 19 (suppl. B)* (2014), 215–236.

[5] Paasonen, P., Nieminen, T., Asmi, E., Manninen, H. E., Petäjä, T., Plass-Dülmer, C., Flentje, H., Birmili, W., Wiedensohler, A., Hõrrak, U., Metzger, A., Hamed, A., Laaksonen, A., Facchini, M. C., Kerminen, V.-M., and Kulmala, M. On the roles of sulphuric acid and low-volatility organic vapours in the initial steps of atmospheric new particle formation. *Atmospheric Chemistry and Physics 10*, 22 (2010), 11223–11242.

[6] Suda, B. Data science salary survey. *O'Reilly* (2017).