Mathematical Institute

Internship Report

Statistical Science for the Life and Behavioural Sciences

Relation between Surface NO_2 Concentration and NO_2 Column Density

Author:
Deirdre I.R. Douma

Internal Supervisor:
Dr. Ankie J.M. Piters
Koninklijk Nederlands Meteorologisch Instituut

External Supervisor:
Prof. Cajo J.F. ter Braak
Wageningen University and Research Centre

July 2016

Chapter 1

Royal Netherlands Meteorological Institute

For my internship I worked at the Royal Netherlands Meteorological Institute (Koninklijk Nederlands Meteorologisch Instituut - KNMI) at the department of Research and Development Satellite Observation (RDSW). KNMI is the research agency of the Ministry for Infrastructure and Environment. About 460 people work at the institute. The institute was established in 1854 and its first director was Professor C.H.D. Buys Ballot, the famous Dutch chemist and meteorologist. It is most famous for its national weather forecasts, which are presented each day in the news and started in 1924, as well as its warnings on extreme weather in the shape of weather alarms. In general, the activities within the institute focus on weather, climate and seismology. As defined in the Law on KNMI, the institute provides

- general weather forecasts for the public sector,
- meteorological information for the aviation sector,
- conducts research and advises the Ministry on meteorological and geophysical phenomenons, and
- maintains the national infrastructure of the meteorological and geophysical fields.

KNMI has thirteen different departments, three of which involve supporting, staff and IT related activities. Four departments focus on research and development, namely: R&D Weather- and Climate Forecasting, R&D Observations and Datatechnology, R&D Seismology and Acoustics, and lastly R&D Satellite Observations (RDSW), where I completed my internship. The other departments focus, in short, on data validation and storage, the weather warning alarms, weather and calamity forecasting, weather and climate services for aviation, and climate change.

The RDSW branch generates and interprets satellite data to obtain a deeper understanding of weather and climate processes. The department is quite large, about 45 people work there, and there is wide range of activities. For instance, apart from interpreting and validating satellite data, the department also develops specifications for new satellite projects, such as the TROPOMI. My first supervisor, Dr. Ankie Piters, specializes in remote sensing of nitrogen dioxide (NO_2), which both involves validating data on NO_2, as well as coordinating international projects and installing new measurement instruments.

KNMI is involved in many international projects and is the organization that translates the findings of the Intergovernmental Panel on Climate Change, World Meteorological Organization, European Union and Climate Summits to the Dutch case. It further advices the Dutch ministries on future courses of action and policy. In addition, it works closely with other organizations and ministries, such as the European Space Agency, SRON Netherlands Institute for Space Research, Rijksinstituut voor Volksgezondheid en Milieu (RIVM) and Planbureau voor de Leefomgeving (PBL) to name a few.

Chapter 2

Activities

My internship assignment revolved around finding a statistical relationship between surface NO_2 concentrations and NO_2 vertical column densities (VCD) for Cabauw, the Netherlands. NO_2 is an indicator for air pollution, which is measured by around 60 weather stations in the Netherlands. However, these stations are not spread around the country evenly; most of them are located within the Randstad. Hence, for remote areas such as Friesland and Drenthe only a few measurements are available.

Another way to obtain NO_2 measurements is by using a satellite. Satellites measure the vertical column densities in the troposphere, which is related to, but not the same as surface concentration. Surface concentration only covers the first couple of meters from the ground, whereas vertical column density reaches from the surface to the top of the troposphere (at least 9 kilometers height) and is measured in another unit (i.e. molecules per cm^2). Every day, the OMI passes over the Netherlands and obtains NO_2 levels for all regions in the country. Hence, by using a satellite a measure for NO_2 concentration for even the most remote places can be obtained. My research goal was to define a model that can predict surface concentrations from these satellite-obtained VCDs using additional meteorological parameters, such as wind speed, humidity and sun duration among others.

I had my own desk at the KNMI, where I shared an office with a programmer involved in the software development of the new NO_2 measurement satellite. Hence, he was able to help with writing the code I needed to fulfill my research. I worked on the project independently and reconvened with my supervisor at the KNMI on a weekly basis to discuss my progress, to obtain additional data and background information when necessary, and to receive feedback. During my internship I had to give two presentations on my research. In addition, I visited multiple other scientists in the institute for specific questions concerning the data or the physical background of the variables I used.

My internship started out with reading and summarizing previous research completed on the same topic. About four similar analyses were performed for other places on the Earth. However, the statistical techniques used in these studies were very basic: Only simple and multivariate linear regressions were used. In addition, I could not find any indication that the data were corrected for autocorrelation, even though it was time-series data. Common procedure within the KNMI when analyzing these kinds of data also involved just basic regression. Further techniques such as linear mixed models and Bayesian methods were unknown or only vaguely familiar.

Hence, one of my main activities was explaining the shortcomings of simple and multivariate regression and the benefits of using more advanced techniques, both to my supervisor as to other scientists that were aware of my activities. See also the presentation which shows how the total variation explained of log(surface NO_2) increases when using a linear mixed model instead of a multivariate linear regression.

I analyzed the relationship between surface and vertical column densities both using in-situ/remotely sensed data as well as using data simulated by the atmospheric computer model EURAD. A linear mixed model was applied with an autoregressive covariance structure to account for the correlation between observations and a random effect for daily variation. In addition, I added two variables to account for diurnal and seasonal patterns in nitrogen dioxide concentration. Differences in the data and the conclusions drawn from both data sets were evaluated in their respective physical background for a one month time window, i.e. May 2012, since only for this time frame data on nitrogen dioxide was simulated by the EURAD model. Re-analysis of the in-situ/remotely sensed data was performed using a larger data set covering nearly 2600 observations taken between March 2011 and December 2012.

Meteorological parameters indicated as important by the literature were included in the models to check for their relevance. Most meteorological parameters were readily available. However, data on atmospheric boundary layer (ABL) height, which is considered to be one of the most important factors, was only available in its unvalidated form. Possible methods of trying to convert the unvalidated data on ABL height into a useful form by hand or algorithm were discussed with colleagues from other departments, but proved too much work for the time I had at KNMI. This was because ABL height can only be measured indirectly and is not measured with great accuracy. Instead, simulated ABL height data from the EURAD model was used. However, because the data from the EURAD model showed some deviation from the in-situ/remotely sensed data, this added uncertainty to the analysis. ABL height was a significant indicator for the data set as simulated by the EURAD data, but not for the in-situ/remotely sensed data.

The physical meaning of the significant meteorological parameters was evaluated. Diurnal and seasonal patterns in NO_2 levels, amount of sunlight, mist and wind speed all influence the relation between surface concentration and vertical column densities of NO_2. Especially the significance of mist was unexpected, since previous research did not indicate it to be relevant. However, during a measurement of the vertical profile of NO_2 using a weather balloon performed by the KNMI in 2010, it was seen that surface concentration was very low compared to NO_2 concentration at ABL height during mist.

The produced models were used on NO_2 vertical column densities obtained by the satellite instrument called OMI and meteorological parameters for Cabauw. The predicted surface concentrations were compared to in-situ measured surface concentrations to see which model gave the best predictions and to obtain a measurement of the accuracy of prediction. The distribution of the random effect daily variation showed a relatively large variance. Since daily variation in NO_2 levels in fact represents undefined background variables (i.e. there is no physical basis that day of the year influences NO_2 concentration), in all likelihood regional effects, further improvement in prediction is possible by trying to incorporate these missing variables.

Before leaving KNMI, I discussed with my supervisor there and some other colleagues this and other further improvements that can be made to the model, e.g. further nesting of the data, switching to Bayesian analyses in R, possible algorithms to convert the unvalidated ABL height data into a useful form, etc. My supervisor was interested in continuing my research in order to reduce the prediction error and so make the model more applicable.

Chapter 3

Skills Acquired or Strengthened

Because the statistical knowledge at the department was not very advanced, I got a lot of training in explaining the statistical techniques into layman's terms. This included basic concepts such as p-values, what it means and what it doesn't mean if something is significant, and the need for assumption checking. The presentations I held were in front of people from all departments of KNMI, some of which only had little experience with data analysis. Hence, I got a lot of questions on the linear mixed model method, and since the discussion rounds after the presentations were quite short, it forced me to explain these concepts as clear as possible in as short a time as possible. At first, I had some difficulty with this. I kept falling back to a point in which I assumed a certain amount of basic knowledge of statistics. But along the way, I got more and more comfortable.

As a helpful tool for my supervisor, I wrote her a short appendix explaining all methods and tests I used during my internship, such as which assumptions needed to be checked for what reasons, how outliers can influence test results, and why main effects should remain included in the model when the interaction effect is significant. All basic stuff, but because I had to go over it time and again, I was really able to internalize these concepts fully. Like Einstein said: "If you can't explain it simply, you don't understand it well enough." The appendix also clearly helped our discussions.

While I was trying to educate my co-workers on the benefits of delving deeper into statistics, they were constantly pushing me to translate my findings into practical physical applications, which was very nice, but also new to me. For instance, if I would log transform my dependent variable, what was I implying of the relationship between surface NO_2 and NO_2 vertical column density? Or, if an interaction term is significant and their main effects aren't, how is this different from the case in which both interaction and main effects are significant in terms of the underlying physical meaning? In order to explain these concepts, I ran a couple of small simulations with very simple data.

KNMI prefers it if all work is done in Python, not R. At first I was happy to have the opportunity to learn how to write code in Python, but after a while I ran into its shortcomings. The code in Python is quite similar to R, and though programming was slow at first, after a couple of weeks I got the hang of it. There are packages available in Python for basic statistical analyses, such as simple and multivariate regression, but anything more complicated (including generalized least squares) demands a lot of programming by hand. This did force me to work efficient and develop some nice functions, as well as return to the actual equations of the methods employed. However, I have fully come to understand the greatness of R. It also made me wonder if this was not one of the reasons why the habit of doing statistical research at the department was at such a basic level.

At the end of my internship I had to summarize all my findings into one report for KNMI, which was subsequently published on their intranet. Unlike writing reports for the courses in our studies, I had to find a mix between general statistical practices in reporting and practices that were common within KNMI and the fields of meteorology and physics. It also involved adding a lot of extra explanations to make the text understandable to people with little statistical background, while at the same time going as deep as I went during my internship. Though I have some experience writing in academic English, having to write directed to an audience with whom I am not completely familiar is always a challenge. Hence, this was a great training.

Even though I have had a couple of jobs, I have never worked in a proper office environment before and it took some getting used to. I shared my office with one other person, who hadn't shared his office in years with anyone. He had the habit of talking to himself while he was working, which demanded some getting used to. Other interns were spread around the building. I only worked four days in the week; I had Monday's off, which was a shame, since each Monday there was a lunch organized for all interns. My desk was also completely on the other side of the building from my supervisor and her direct colleagues which also worked on the same data. Hence, networking required a bit of an effort from my part, especially since I am not a coffee drinker and the coffee machine seemed to be the place where things were happening.

The really nice thing about working at KNMI is, however, that there are a lot of presentations given on a variety of topics from all the different departments each week. Hence, I got the opportunity to learn about the model used to forecast the weather, whether or not climate change played a role in the conflict in Syria, etc. I also received some great feedback for my own work and people were eager to learn new things and talk in length about their own research and findings. Other researchers there came from a variety of different background and places, from physicists, programmers, to historians, and the Netherlands, Jordan to China. They also cared about improving their knowledge. For instance, my supervisor may want to continue my research by finding a new intern who will convert my findings into a publishable article.

Chapter 4

Conclusion

My internship at KNMI at the department of Satellite Observations was very educational to me. It gave me the opportunity to work at the most leading institute on climate and meteorology in the Netherlands and combine both my interest in statistics and climatology. It was exactly this combination of the two fields that drew me towards the KNMI in the first place. During the internship I both had to delve into the physical world of how nitrogen dioxide behaves in our atmosphere as well as the statistics involved in how to model and predict time-series data, and delve into how to combine these two different areas.

The internship allowed me to experience what it is like to be a statistician at an institute where a lot of different types of research takes place, and where everyone has its own field of expertise. I was the first trained statistician at the department, and so once word got out, I had people drop by asking me all kinds of questions on statistics and how they should apply it or what other possible methods there were for their own research. I sometimes felt like I was becoming the go-to-girl when it come to statistics in the department, which was a great honor, though at first also a bit daunting.

It also gave me the opportunity to further develop myself as a programmer. I become experienced in working with Python and along the way created a bunch of neat functions to evaluate the data and construct plots for my mixed models. I am actually considering of making them a bit more generic and combining them as a Python package. Furthermore, I had the opportunity to fully explore LaTeX and the interface between Pyhton and LaTeX, as well as further develop my academic writing skills.

In addition, it was very good to work in an environment for a while where not everyone knew straight away exactly what I was doing. This really forced me to take a couple of steps back and return to the core of what I was doing. Even though I became much better at explaining the statistical methods and procedures in simple terms, I still struggled sometimes, especially when talking to someone who was a hard-core physicist who had been doing his or her research for years in a certain way.

Working at KNMI also introduced me to office politics and how to behave within it. Thanks to this internship I now know that I feel comfortable in playing the role of statistician within a larger research framework. I feel more motivated then ever to become a fully functioning applied statistician that can help an institute such as KNMI to fulfill an important research task in society.