How’s the Air Out There? Using a National Air Quality Database to Introduce First Year Students to the Fundamentals of Data Analysis

David Hall and Jessica D’eon (Corresponding Author)

19/03/2021

# 1 Abstract

# 2 Introduction

Whether we like it or not we are living in an age of data, and the world of chemistry is no exception. From big-data atmospheric chemistry in climate-change models1 to machine-learning organic synthesis2, every domain of chemistry is increasingly relying upon data-driven science. In response to this trend, undergraduate chemistry curricula need to adapt to better equip and prepare the next-generation of chemists with the necessary skills and knowledge to navigate an increasingly data-centric world. As we ourselves work on new undergraduate teaching material, we notice that an oft-overlooked aspect in how data analysis is presently tough is how exactly data (measurements, signals, etc.) is transformed into information (trends, correlation) and finally into knowledge. The explicit teaching of these concepts is often neglected in current teaching labs, resulting in increasing student frustration. Motivated by this, and the need to transfer to a virtual laboratory environment as a result of Covid-19 social distancing restrictions, we sough to develop a new, remote learning compatible, experiment.

An obvious assumption of teaching data science is that students will eventually analyze *real* data, which is often permeated with outliers and the fingerprints of gross experimental errors. The data that undergraduate chemist will collect in their academic careers is no different. However, acquiring sufficient data for analysis if often stressful for undergraduate students in a teaching laboratory setting due to time- and equipment-constraints. As well, with the Covid-19 restrictions students were unable to attend labs, and hence unable to produce any of their own data. We saw this as an opportunity to integrate actual measurements from published data repositories. A choice example is the air quality data from Environment and Climate Change Canada’s national Air Pollution Surveillance Programs (NAPS). There is no shortage of data to be analyzed, because the NAPS program has been conducting hourly measurements across Canada since 1975 of several major atmospheric.

Prominent atmospheric pollutants are structurally simple, and undergo reaction schemes comparable to those covered in introductory chemistry lectures. Ozone (O3) and nitrogen dioxide (NO2) are two choice candidates for analysis by undergraduate students. They are structurally simple molecules, and undergo reaction schemes comparable to those covered in introductory chemistry lectures (see reactions (2.1), (2.2), and (2.3)). Notable of these compounds is their interdependent diurnal cycles. The relationship between O3 and NO2 is so intimate, the term “odd-oxygen” (Ox) is used to express the sum of these two compounds (see reaction (2.4), and ref.),3 although the relationship between O3 and NO2 can vary with environmental and anthropogenic influences including temperature, sunlight, and motor vehicle emissions.

While students investigate seasonal differences in the relationship between atmospheric O3 and NO2 from subsets of the NAPS dataset, our Air Quality lab explicitly introduces data analysis work-flows. As students work through the instructions they are encouraged to generate hypothesis through probing questions, which they further investigate across the entire NAPS dataset through a new interactive application we developed. As the entire lab uses previously acquired data, and the ubiquitous Microsoft Excel software package, students were able to explore real data from home, and complying with Covid-19 restrictions.

# 3 Experimental Overview And Pedagogical Goals

This 3 hr data-analysis laboratory exercise uses publicly available data and open-source code (described in the [Supplementary information]), and has been run successfully in the one-semester “CHM135: Chemistry: Physical Principles” undergraduate general chemistry at the University of Toronto since Summer 2020. This course is most often conducted in the first-term of the first-year of life-sciences/chemistry undergraduate students. As our *Air Quality Lab* being the first-lab of five, it is designed as much as an tutorial on data-analysis and Microsoft excel as it is to explore atmospheric chemistry. The lab is divided into three parts: the prelab, data analysis in Excel, and data exploration & hypothesis generation.

The prelab follows a traditional approach, and is written to situate students in the relevant chemistry for the upcoming analysis. Specifically for this lab we create explanatory videos and material introducing gas phase chemistry, and relating the lab content to concurrent lecture material of gas phase chemistry (i.e. ideal gas law).

In the data analysis portion of the lab students are randomly assigned two datasets. Each dataset is a 7-day snapshots of hourly O3 and NO2 measurements taken from the NAPS program. The datasets are all from the same NAPS surveillance station for a given year. For our purposes we chose a different downtown Toronto NAPS station for each successive iteration of the lab. The two datasets correspond to 7-days in the winter and 7-days in the summer, and were generated from original NAPS data using R as described below. Alongside their data, students are provided with a written handbook detailing the necessary Excel operations, and an synchronous online session with their TA. Working through the lab exercises students are explicitly taught data analysis workflows, modeled after that recommended by Hadley and Grolemund4:

1. *Importing* their assigned comma serrated values (.csv) data sets into Excel.
2. *Tidying* their data and setting up their worksheets. This step consists of formatting cells to properly display values and handling missing data. Specifically for this lab, the NAPS dataset stores missing values as ‘-999,’ which can be erroneously interpreted literally by Excel.
3. *Visualizing* their data by creating a time-series plot of time vs. concentration of each pollutant, see Figure 3.1.
4. *Transforming* their data using mathematical operators in Excel to calculate total oxidant and adding it to their time-series plot as well as calculating 8hr moving averages.
5. *Modeling* a linear relationship between [O3] and [NO2] to qualitatively assess the negative relationship between these two contaminants.
6. *Communicating* and exploring their results through a series of accompanying questions.

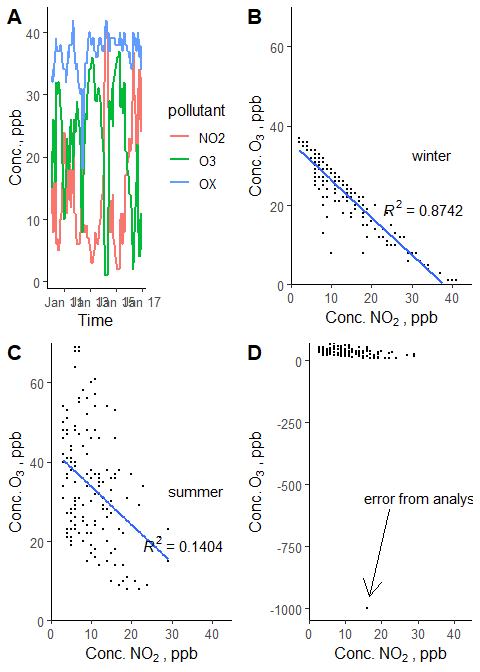
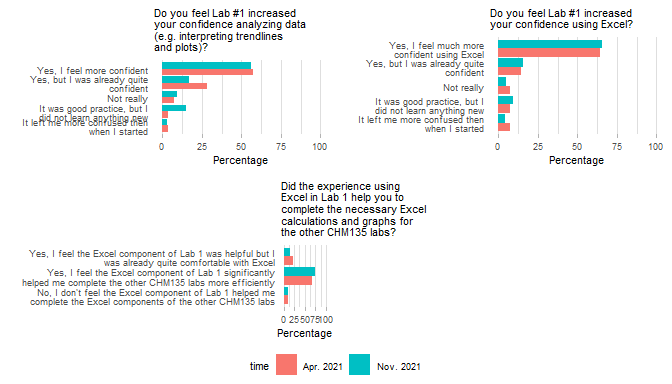


Figure 3.1: Example of plots students are expected to create. (A) time-series of pollutants across 7 winter days. (B) Correlation plot of O3 and NO2 concentrations with linear regression in the winter and (C) summer data sets. (D) Example plot if a ‘-999’ value wasn’t removed.

The last step in this workflow is expanded in the final part of the lab where students compare their results to the complete NAPS dataset from which their assigned datasets originated from. Here they are encouraged to generate hypotheses based on their own data, and their *a priori* chemical knowledge introduced in the prelab. To this end, we created an interactive online application that students visit using *R* and *Shiny*5. This application consist of an interactive map showing the location, and local population, of every NAPS surveillance station contained in the dataset. Students can then select any station and time-span, and a time-series and correlation plot, similar to the ones they created themselves, are automatically generated. This allows them to rapidly compare their data to any number of stations, simultaneously relieving them of the burden of repetitive and tedious data analysis while facilitating hypothesis generation and data exploration. Accompanying questions prompt students to explore and reason differences in O3 and NO2 correlation between urban and rural areas, as well as between winter and summer datasets. See the [Supporting Information](#supporting-information) or <https://davidrosshall.shinyapps.io/AirQualityApp/> for details on the application.

# 4 Results and Pedagogical Outcomes



Based on the our surveys and feedback from students, the majority of students responded positively to the new learning experience. From our survey of students in the most recent iteration of the Air Quality Lab in the winter 2021 term (n = ??), 68% of respondents stated they felt the Excel component of Lab 1 significantly helped them complete the other CHM135 labs more efficiently. (Figure ??). This is a welcomed improvement as students frequently complain about the time commitment required for the CHM135 lab component. Students also feel more confident with regards to overall interpretation of data (plots, trendlines) as well as towards their use of Excel in general (57%, and 64%, respectively) (Figure ??).

Included in the survey was the option for students to provide any additional feedback on the lab. Students expressed both positive and negative feedback to the Air Quality lab (See [Supporting Information](#supporting-information) Table ?? for complete feedback). Students appreciated the introduction to Excel, the practical usefulness of the incorporated material, and the opportunity to analyze real world data offering a glimpse into environmental chemistry. However, students were also critical of how the material was implemented. Some experienced trouble with inconsistencies between the Excel instructions, and their version of Excel (although all UofT students are provided with free access to the latest version of Excel, with guidance/links provided in the aforementioned lab instructions document). Likewise, many students felt the Excel component should have been explicitly tough during the synchronous session, rather than those sessions focusing explicitly on lab material (i.e. data analysis vs. Excel operations). The CHM135 contains hundreds of students, many of whom have prior experience with Excel. Consequently, we opted not to directly teach students the basic workings of Excel in the synchronous session, as this would have bored many of students. Going forward, we feel that incorporating optional Excel help-sessions specifically to assist students with this component of the lab.

A major addition to this lab is the development of the Air Quality App using Shiny. We strongly believe this was a great component, as the functionality of the app greatly reduces the ‘friction’ allowing students to readily explore the larger NAPS dataset without burdening them with lengthy data prep/analysis. Furthermore, as students explore the larger NAPS dataset they see data ‘unknown’ to the instructors. As the data is from real and complex environments, students often find data that *does not* support their hypothesis or that contradicts the pre-lab information! A stark contrast to the traditional undergraduate lab which is tailored so that students produce predictable results. The framework of the latter inadvertently teaches students to fit their data to the theory, in opposition to actual research where the data drives the theories. In other words, students are excited to make real connections and real discoveries with real data rather then going through the often tedious motions of analyzing pre-screened or synthetic data.

# 5 Conclusions

We sought to create a new introductory lab experiment to expressly teach incoming students fundamental components of data analysis/science as well as practical instructions on how to use Microsoft Excel to for a solid foundation for future courses. To this end, we leveraged the R computer language for the automated, and scalable, generation of unique data from real world atmospheric measurements from the NAPS that served as the basis of our introductory data science lab. Alongside written instructions on how to use Excel, we developed an online interactive App, allowing students to readily explore the entire NAPS dataset to compliment their individual analyses. Our efforts were rewarded with students being better equipped to tackle subsequent data analysis challenges in the following introductory chemistry labs, in addition to arming them with skills they will assuredly make use of outside of the first-year chemistry laboratory.

# 6 Supporting Information

# 7 Author Information

David Hall, Department of Chemistry and School of the Environment, University of Toronto.

Jessica D’eon, Department of Chemistry and School of the Environment, University of Toronto.

# 8 Acknowledgements

# References

(1) *The Future of Atmospheric Chemistry Research: Remembering Yesterday, Understanding Today, Anticipating Tomorrow*; National Academies of Sciences, Engineering, and Medicine (U.S.), Ed.; The National Academies Press: Washington, DC, 2016.

(2) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nature Reviews Chemistry* **2019**, *3* (10), 589–604. <https://doi.org/10.1038/s41570-019-0124-0>.

(3) Kley, D.; Geiss, H.; Mohnen, V. A. Tropospheric Ozone at Elevated Sites and Precursor Emissions in the United States and Europe. *Atmospheric Environment* **1994**, *28* (1), 149–158. <https://doi.org/10.1016/1352-2310(94)90030-2>.

(4) Wickham, H.; Grolemund, G. *R for Data Science*; Beaugureau, M., Loukides, M., Eds.; O’Reilly Media Inc.: Sebastopol, CA, 2017.

(5) Chang, W.; Cheng, J.; Allaire, J. J.; Xie, Y.; McPherson, J. *Shiny: Web Application Framework for r*; 2020.