CTFP Final Report

David Hall and Dr. J. D'eon (supervisor)

2020-09-04

Introduction

Whether you like it or not, we are living in an increasingly data centric world, and the field of chemistry is no exception. An oft overlooked aspect of this is how exactly data (measurements, signals, etc) is transformed into information (trends, correlation) and finally into knowledge. Moreover, the explicit teaching of these concepts is often neglected resulting in increasing student frustration.(2) Motivated by this, and the need to transfer to a virtual laboratory environment as a result of Covid-19, we saught to develop a new, remote learning compatible, experiment for *CHM 135: Physical Principles*.

Experiment 1: The Chemistry of Air Quality is the results of our efforts. In this new experiment first-year students are introduced to fundamental data analysis concepts as they explore some of the chemistry of airborn pollutants.

Chemistry background

Since 1975 Environment and Climate Change Canada (ECCC) has been monitoring several airborn pollutants through the National Airborne Pollutant Surveillance (NAPS) program. Two of the key pollutants monitored are ozone (O_3) and nitrogen dioxide (NO_2), and whose interdependent diurnal cycles are expressed through equations 1 to 3, right. The relationship between O_3 and NO_2 is so intimate, atmospheric chemist have developed the tern "odd oxygen", O_x, as the sum of these two components (equation 4).(1) Lastly, the correlation between O_3 and NO_2 varies with environmetal influences such as increased levels of volatile organic compounds and temperature engendered during the summer months. Through the NAPS data, students can visualize and qualitatively assess these relationships.

$$NO_2 + photon \rightarrow NO + O \quad (1)$$

$$O + O_2 \rightarrow O_3 \quad (2)$$

$$O_3 + NO \rightarrow O_2 + NO_2 \quad (3)$$

$$[O_3] + [NO_2] = [O_x] \quad (4)$$

Experiment workflow

Operationally, after an introductory pre-lab prepared by Dr. D'eon (with accompanying video and gas phase chemistry questions), each student analyzes one randomly assigned winter and summer data sets from a pool upload and distributed through Quercus. Each dataset comprises a 7-day snapshot of O_3 and NO_2 concentrations as measured by a downtown Toronto monitoring station from the NAPS program.[1] The experiment instructions, and a supporting *Tip Sheet* on operations in Excel necessary for Experiment 1, guide students through the data analysis workflow made popular by Wickham and Grolemund.(3)

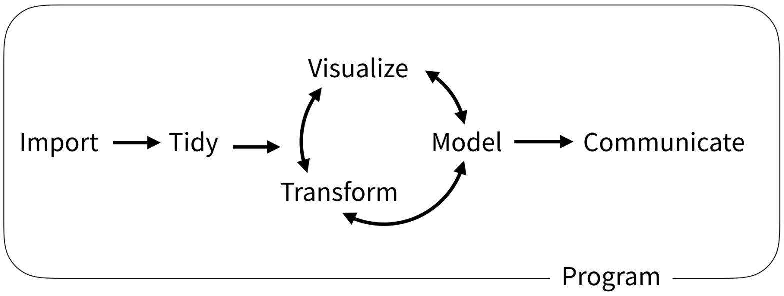

Figure 1: Data exploration/analysis workflow; figure from *R for Data Science* (2017)

- *Importing* their assigned comma separate values (.csv) data sets into Excel.
- *Tidying* their data and setting up their worksheets. This step consist of formatting cells to properly display values and handling missing data.[2]
- *Visualizing* their quantitative information through a time-series plot of time vs. concentration of pollutant.
- *Transforming* their data using mathematical operators in Excel to calculate total oxidant and adding it to their time-series plot as well as calculating 8 hr moving averages.
- *Modelling* a linear relationship between O_3 and NO_2 to qualitatively assess the inversal relationship between these two contaminants.[3]
- *Communicating* and exploring their results through a series of accompanying questions written by Dr. J. D'eon.

Expected student outcomes

Beyond the introduction to gas-phase and atmospheric chemistry, students are expected to learn the basics of data analysis and operations in Microsoft Excel. through the Experiment and supporting *Tip Sheet*. Topics covered include: cell referencing, mathematical operators, *find and replace* functions, cell formating, plotting, and summary statistics. While only touching the surface of data analysis, we believe this list touches upon the most frequently used operations in Excel, and provides a solid base from which students can improve their understanding and skills on their own or in future classes.

[1] A more technical explanation of how NAPS data was subseted for students can be found at page 4.

[2] Specifically, NAPS stores missing values as -999, but this value is literally interpreted by Excel, requiring removal before further data visualization/analysis.

[3] This is accomplished using the "add trend line" function in Excel, although previous versions of the lab utilized the "linear regression" function of the *Analysis Toolpak*.

Verification of student learning is assessed through through the results of each students data analysis. For each data set, students are expected to plot a time-series of polutant concentration and a correlation plot of O_3 and NO_2 with linear regression (Figure 2A and 2B). Additionnaly they perform the same analysis on both the winter and summer datasets, illustrating the increasing complexity of summer vs. winter atmospheric chemistry (Figure 2C). Through their visulizations, students must answer a series of accompanying questions wherein they inquire about possible explanations for the differences in their winter and summer results. Any faults in a student's data analysis are readily aparent in their visualations. For example, an errant '-999' value leftover from the NAPS dataset is easily visible to the TA (Figure 2D). Lastly, every student dataset that is generated is accompanied with by a PDF anwser sheet containing a time-series plot, correlation plots, and all summary statistics students are expedcted to perform. Consequently TAs can simply compare a students analysis of a given datasets against the accompanying anwser sheet to quickly verify their work. How the answer sheets are generated is discussed below.

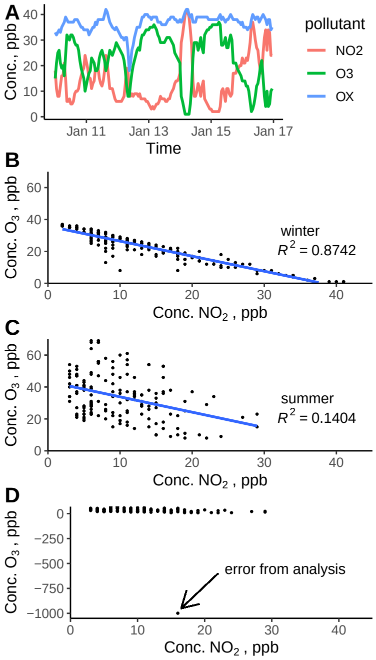

Figure 2: Example of plots students are expected to create. (A) time-series of pollutants across 7 winter days. (B) Correlation plot of O3 and NO2 concentrations with linear regression in the winter and (C) summer. (D) Example plot if a '-999' value wasn't removed.

Lab Results

We were unable to introduce a survey to undergradaute students that would have explicitly addressed their experinences and thoughs with Experiment 1. However, discussing Experiment 1 with the four laboratory TAs revealed two promising insights. Firstly, the TAs found that students' questions were largely related to lab content, and not the technical aspects of Excel. Secondly, student questions that did pertain to Excel were readily addressed by directing them to the supporting *Tip sheet* document.

Furthermore, after inspection of more than 300 student submitted figures, none had any critical flaws (either in visualization or data analysis). There were instances of several minor issues common to several student plots (i.e. adding a linear regression to their time-series plots), but these have been addressed in the updated Experiment 1 instructions and supporting *Tip Sheet* for the Fall 2020 session.

Implementation in the Fall Term

Experiment 1 is slated to be introduced in the upcoming Fall 2020 CHM 135 session, with an estimated >1500 enrolled students. Fortunately, the scope of the NAPS program and the readily scalable coding used to generate student datasets/answer sheets can easily cope with the increase in course participants. As well, a limitation of the summer session was that limitations in the Quercus setup practically limited the entire course to 15 winter and 15 summer datasets. This has been addressed in the upcoming session where each lab section will have it's own unique pool of datasets from a unique NAPS monitoring station in Toronto. While we'd ideally like to assign each student with a unique dataset, this method best minimizes overlap in student analysis while remaining logistically practical within the existing Quercus infrastructure.

Future Directions & Personal Speculations

If someone understands Excel, the actual laboratory exercise is trivial and can be easily completed in half an hour. I think that is perfectly alright, as the principle aim of Experiment 1 is to assure all incoming students have a common-core understanding of data analysis and operations in Excel.

- Dialing in Quercus setup to expand course components to > 1800 students.
  - can create that many data sets/answer keys easily, bottleneck is upload to quercus.
- refine Excel operations
- Explicit discussions on data analysis
  - this lab, in my opinion, is more about understanding data analysis than it is chemistry. Once you learn stuff here you can apply in all sorts of ways in upper year courses.
  - That's why the *Tip Sheet* i wrote was so long. I never expected students to read the entire thing, but if they had any questions they could look it up there. To be fair though, I think the info within that document should prettied up and set up as a departmental wide guide to data visualization/analysis etc. Some of the stuff made by grad students is awful/deceptive.
- Enhanced discussion on statistics with a focus on interpreting the numbers rather then calculating them with mathematical formulas.

Stuff left on the cutting room floor

A surprisingily difficult aspect of creating Experiment 1 was deciding what *not to* include in the course. While we're largely satisfied with the present course, some of the elements could prove useful for future interations or upper year classes. These ideas can be readily re-integrated into the existing experimental framework if needed. Some of these ideas are briefly described below. We hope that they inspire readers for future versions of this lab.

- $S0_2$ is the main pollutant responsible for acid rain, and was regualted nationally in 1985 and provincially in 2007. Consequently, we wanted students to investigate if SO_2 levels have significantly decreased as a result of these regulations. NAPS data of SO_2 emissions cover more than 50 statiosn in Ontario and coveraged extends as far back as 1975. While a great oppertunity to discuss statisical significance, and readily executable in Excel, we struggled with exploring the larger chemistry of this phenomeone in a context approriate to the CHM 135 course.
- The current lab utilizes the "add trendline" option in Excel. We had original envisioned using the *Analysis Toolpak* which extends the linear regression beyond a line of best fit and R^2 to significance of the coefficeincts (slope) and ANOVA analysis. While the greater exploration of linear regression was promising, the stastical discussionw as determined to be beyond first-year chemistry students.
- Building off of the last point, we had integrated historical ECCC weather data E[4] into the analysis. We envisioned students comparing the correlation of O_3 with NO_2 and with temperature (or other weather phenomena). This was abandonned for the present winter vs. summer.
  - There exist an oppertunity to expand this component to multiple linear regression utilizing pollutant concentration and various weather readings, although this is more appropraite for an upper year environmental chemistry course.
- Since O_3 and NO_2 levels are largely affected by anthropogenic activity, we wanted to analyze the effects of the Covid-9 related "anthropause" on downtown Toronto levels. Unfortunatly the NAPS datasets has a one year lag time to publication, and the live feed from the Ontario Ministry of the Environment[5] requires either significant data collection from students or web-scrapping. The former is unreasonable as the entire course could be derailed if the website goes down, and the later is beyond the skills/contract hours of this humble TA.
  - This may be worth revisiting in upcoming years when the 2020 NAPS dataset is realeased.
- The NAPS program spans the entirety of Canada. We explored comparisons between rural and urban NAPS mesurements but found the process quickly excalated beyond a CHM 135 appropraite discussion.

[4] Hourly weather data downloadble here:`https://toronto.weatherstats.ca/download.html`

[5] MOE real time air quality data can be found here: `http://www.airqualityontario.com/history/summary.php`

Source code and instructions for generating datasets

Generation of all of the data sets and their accompanying answer sheets are executed in R with prodigious use of the *tidyverse* and *Rmarkdown*. As it stands we can easily generate a unique pairing of summer/winter datasets for evevery student in the Fall 2020 session. The source code and example ECCC data, student datasets, and TA answer reports can all be found on GitHub.[6]

A more detailed description of the process is described in the GitHub repo, but briefly:

1. ECCC hourly data for NO_2 or O_3 is subset based on a given reporting station; in the present iteration all measurements are from downtown Toronto in 2018.
2. Bilingual headers, ancillary columns, etc. are removed from the ECCC dataset, dates are conversed to Excel format, and the remaining data is transformed from the 'wide' matrix style to the 'long' columnar format for easier manipulation in Excel, see table 1.
3. A specified number of student datasets are generated from a 7-day moving window of the year-long data. I.e. dataset 1 is January 1st to 7th, dataset 2 is January 2nd to January 8th. A complimentary summer dataset is taken starting from July 1st.
4. A "-999" error is inserted randomly into each student dataset.
5. Datasets saved in a new folder as .csv files.
6. For each dataset generated, a Rmarkdown script generates a PDF with the analysis results. TAs can compare the answer sheet to student submissions.

Table 1: Student data set

Date	NO2	O3
43110.00	18	15
43110.04	11	21
43110.08	8	-999
⋮	⋮	⋮

Again, all of this code is explained on the GitHub repo. For the unawares, GitHub provides hosting for software developement, distribution version control, and source code management and is readily integretated into the RStudio environment. In practice, this means that the code used to automatically genereate student datasets and anwser keys is preserved online, and can be safely passed along from year to year thanks to version control. The GitHub environment is ideal for introducing new componenets and removing old ones from the code thanks to version control, and provides an effective frameowrk to ensure Experiment 1 can be readily changed in its future iterations.

[6] Github repository with source code: `https://github.com/DavidRossHall/CHM135_Exp1Data`

References

[1] Dieter Kley, Heiner Geiss, and Volker A. Mohnen. Tropospheric ozone at elevated sites and precursor emissions in the United

States and Europe. *Atmospheric Environment*, 28(1):149–158, jan 1994. ISSN 13522310. doi: 10.1016/1352-2310(94)90030-2.

[2] Nicholas E. Schlotter. A statistics curriculum for the undergraduate chemistry major. *Journal of Chemical Education*, 90(1): 51–55, jan 2013. ISSN 00219584. doi: 10.1021/ed300334e. URL `https://pubs.acs.org/sharingguidelines`.

[3] Hadley Wickham and Garrett Grolemund. *R for Data Science*. O'Reilly Media Inc., Sebastopol, CA, 2017. ISBN 978-1-491-31039-9.