CTFP Final Report

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

2020-08-28

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 concepsts 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 experiment for *CHM 135: Physical Principles.*

Experiment 1: The Chemistry of Air Quality is the results of these 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 (Needs to be cleaned up)

Since 1975 Environment and Climate Change Canada (ECCC) has been monitoring several airborn pollutants through the National Aurborn Pollutant Surveillance (NAPS) program. Two of the key pollutants monitored are ozone (O_3) and nitrogen dioxide (NO_2), and whose interdepentant dirunal cycles are expressed through equations 1 to 3, right. With the provided datasets students are able to visualize this relationship in a time-series plot as well as qualitatively assess this relationship (see below). Lastly the relationship between O_3 and NO_2 is so intimate, atmospheric chemist have developped the tern "odd oxygen", O_x, as the sum of these two components (equation 4).(1)

$$NO_2 + photon \rightarrow NO + O \tag{1}$$

$$O + O_2 \rightarrow O_3 \tag{2}$$

$$O_3 + NO \rightarrow O_2 + NO_2 \tag{3}$$

$$[O_3] + [NO_2] = [O_x] \tag{4}$$

Experiment workflow

Operationally, each student analyzes one randomly assigned winter and summer dataset, each comprising a 7-day snapshot of O_3 and NO_2 concentrations as measured by a monitoring station from the NAPS program. The experiment instructions guide students through the data analysis workflow made populare by Wickham and Grolemund.(3)

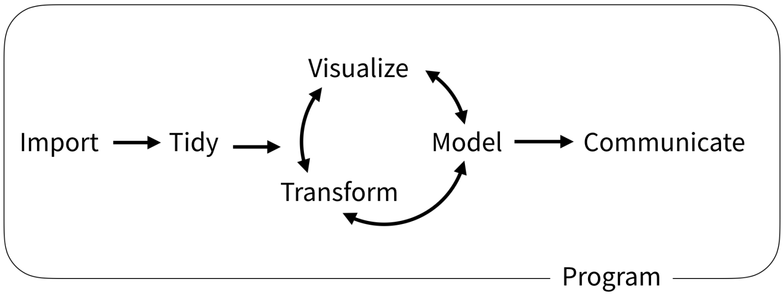

Figure 1: Data exploration/analysis workflow; figure from *R for Data Science* (2017)

- *Importing* their assigned comma seperate values (.csv) datasets into Excel.
- *Tidying* their data and setting up their worksheets. This step consist of formatting cells to properly display values and handling missing data.[1]
- *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.
- *Modeling* a linear relationship between O_3 and NO_2 to qualitatively assess the inversal relationship between these two contaminants. [2]
- *Communicating* and exploring their results through a series of accompanying questions written by Dr. J. D'eon.

Expected student outcomes

- generate plots like figure 2.
- discuss difference between winter and summer datasets (figure 2b and 2c)
- excel functionality (maybe list)

Lab Results

- couldn't really get a survey in
- conversation with TAs revealed some minor issues around Quercus layout and errors in supplementary tip-sheet, these were addressed for upcoming Fall 2020 term.
- Review student submitted work showed most of them understood what was goign on, although a handful of errors (i.e. applying linear regression to time series data.)

[1] Specfically, missing data is stored as -999 in NAPS datasets but this value is litorally interpreted by Excel, confounding data visualization/analysis.

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

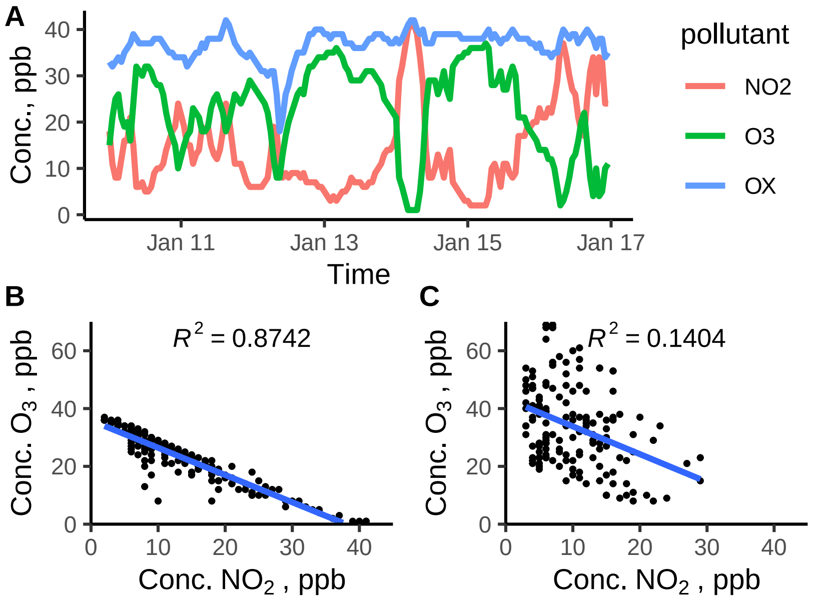

Figure 2: Example plots students are expected to create. (A) time-series of pollutants across 7 winter days. (B) Correlation plot of O3 and NO2 concentrations with linear regression in the winter and (C) summer.

Troubles with implementation

Only uploaded 15 winter and 15 summer datasets.

Future directions

- Dialing in Quercus setup to expand course componenets to > 1800 students.
  - can create that many datasets/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

- tried a bunch of stuff, and left plenty on the cutting room floor
  - SO2 work
  - using Analysis Toolpak for linear regression (outputs additional parameters hidden from display line of best fit)

Source code and instructions for generating datasets

The source code and example ECCC data, student datasets, and TA answer reports can all be found on GitHub.[3]

NEED to explain GitHub better I feel like this is goign to be lost on folks, when it actually solves so many issues about the way information is passed along between faculty and over the years.

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 safely passed along from year to year. The GitHub environment is ideal for introducing new componenets and removing old ones from the code thanks to version control. This is expecially important as faculty frequently rotate through the CHM 135 course.

Brief discussion on how datasets are generated, what transformations need to be applied, etc.

Table 1: A tibble of a student assigned dataset; note the Excel complient data & time formats.

Date	NO2	O3
43110.00	18	15
43110.04	11	21
43110.08	8	25
43110.12	8	26
43110.17	12	21
43110.21	16	19

References

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

[3] Github link: `https://github.com/DavidRossHall/CHM135_Exp1Data`

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.