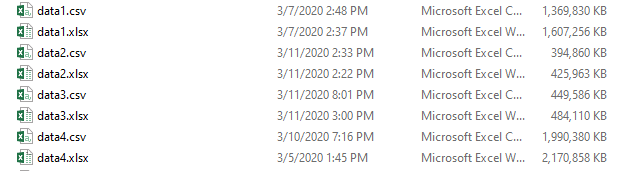
**Extraction, Transformation, and Simplification of High Mass- and Time-Resolution Data**

*Data Specifics*

When sampling online with a high-resolution time of flight mass spectrometer (HR-TOF-MS), we acquire three-dimensional data, with time on the x-axis and mass-to-charge ratio (m/z) on the y-axis. Mass-to-charge ratio refers to the mass of the ionized molecule over its charge. For most molecules and ionization schemes, this charge is +1, so the mass ends up being the mass of the parent molecule plus one ([M+H]+). As opposed to hard ionization, where a stream of electrons cause whole-scale fragmentation and ionization of target molecules, in soft ionization, the goal is to gently impart a charge to the target molecule and reduce the likelihood of fragmentation.

The Vocus PTR-TOF-MS (TOFWERK) can collect data at 1 Hz and can cover an m/z range from 30 Th (Da/e) to more than 500 Th. Lower masses have sensitivity issues and can include interferences from air (m/z 28 from nitrogen, m/z 32 from oxygen, m/z 18, 19, 37, 55, 73, … from water) while higher masses show reduced sensitivity due to variation in ionization and transmission efficiencies. Because of the resolution power of the TOF-MS, we can report m/z’s down to four decimal places.

Though it would seem like this mass spectrometer would yield an unwieldy amount of data, the discrete nature of chemical formulas, with only certain likely chemical formulas, makes the number of masses for the instrument (and for us to analyze) more reasonable. The instrument does check all masses, but it automatically low-pass filters those m/z’s that have minimal signal. Even so, for a series of Vocus experiments of tobacco smoke, the data acquisition software found over 2000 m/z’s with non-trivial values. Data file sizes ranged from approximately 400MB to 2GB for a total of 4.7GB of data in less than a month of sampling on this one instrument. Because we aim to study the data with more granularity and some of these compounds are more important than others, we need to decide which time series data points were worth keeping.



*Data Exporting*

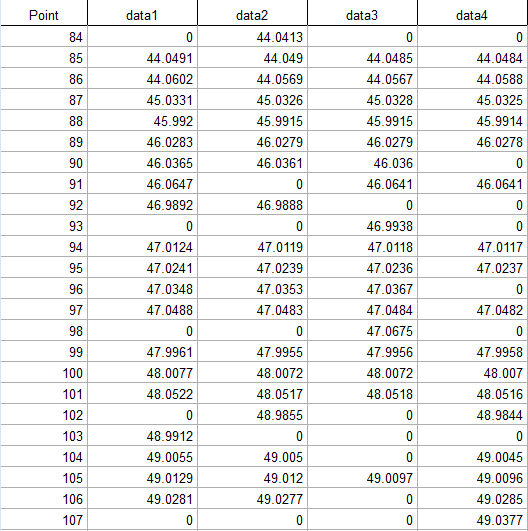
Exporting data from the Vocus is reasonably simple. Data are contained in h5 files, which are a standard format in the atmospheric chemistry field. From there, they can be moved into individual arrays or into the equivalent of a data frame, whichever is easier for future transformations. Our collaborators had exported these data into XLSX files, which I then imported into Igor Pro waves (Igor’s version of 1D arrays), with each wave containing data on each m/z.

*Combining Data Files*

My first task was to reconcile six data sets that cover different stretches of experimentation (12/4-12/6, 12/9-12/12, 12/13-12/15, 12/16-12/20, and 12/20-end). Of these, the 12/13-12/15 and 12/16-12/20 sets contained the most important data. Even if some of these m/z’s theoretically represent the same compound, each data set had slightly different m/z values based on slight differences in mass calibration from data set to data set.

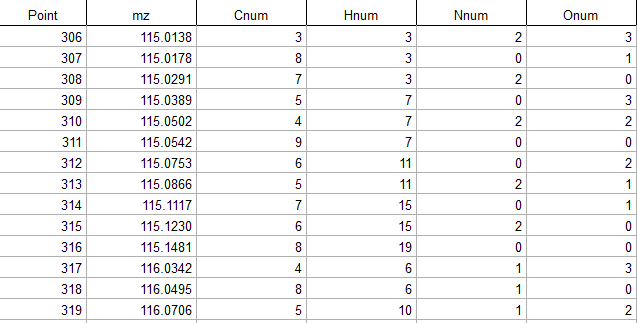
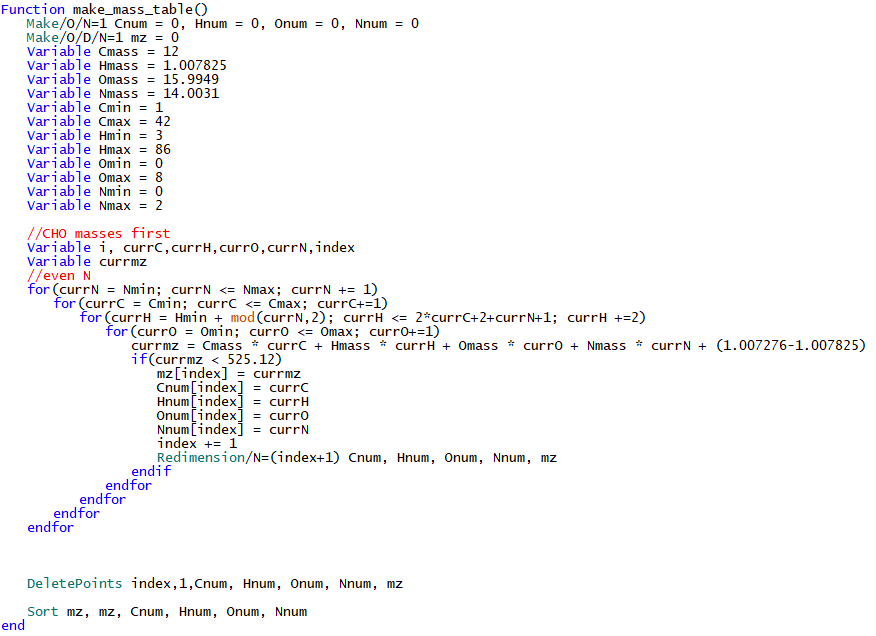
I wrote some code that checked how similar in value these m/z’s were and grouped them if they were sufficiently similar. Before running this code, I combined all the m/z from the four data files, sorted them ascending, and made sure I had a key array to point back to the data file.

Once a given m/z was chosen, I set a maximum resolution for comparison (given by resmax). Imagine this resolution being a window around the target m/z. I then iterated down the list (given by the Wave called mass) and if the subsequent mass was within this resmax, then I grouped them. These were moved to columns named “data#” based on their original data file.

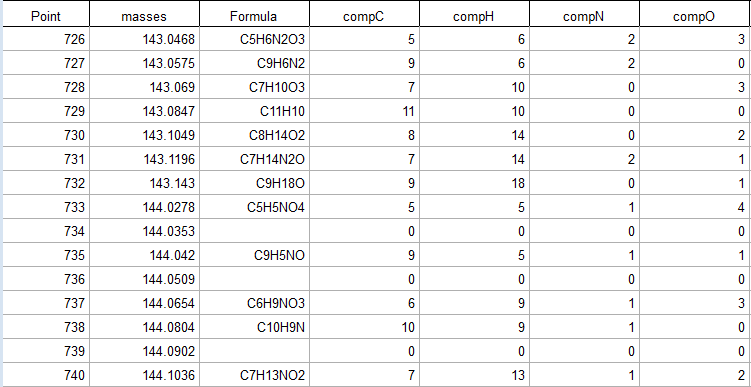
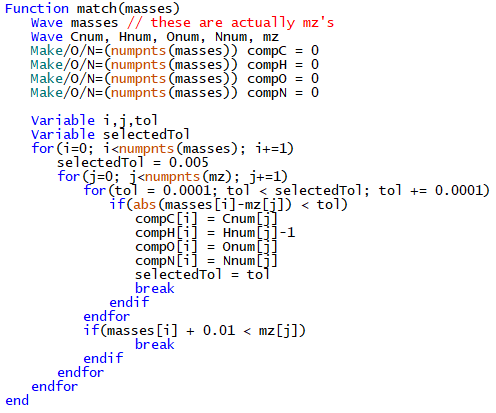


I then did some quick manual checks, especially at higher m/z’s, to ensure that the quick algorithm was making reasonable assignments. Earlier iterations used a fixed max resolution (instead of resmax = mass / 18000 or 0.01) and had trouble at lower molecular weights.

Then, I generated a list (below, right) of all the m/z’s expected for common organic compounds with relatively few nitrogen and/or oxygen atoms (CHN0-2O0-8) (below, left). The nested loops just iterate through the possible C, H, O, and N atom counts, calculating the molecular mass based on their known atomic masses.



Another function I wrote (below left) matched these expected m/z’s with the masses we found in our samples. Yet another function I wrote matched chemical formulas (results, below right) corresponding to average m/z’s for our collection of over 2000 time series.



*Identifying Compounds Based on Their Masses*

Through some previous work I have done with analyzing PTR-TOF-MS data in a German movie theater, I had developed a collection of matches between compound formulas and likely compound identifications, specifically for tobacco smoke data. Using this small database, along with other literature identifications (Koss 2018), I generated compound identifications for as many of these signals as possible (~350) (Main Library of PTR IDs v2.xlsx). In some instances, the formulas had matching identifications upon an Internet or literature search, but insufficient evidence for me to expect them to be present in tobacco smoke. Other times, the presence of isomers (multiple compounds with different structures but the same formula) made accurate identification more challenging.

*Making a Shortlist of Compounds to Focus on*

However, the data set still contained too many m/z’s that were not particularly important. I first took the data specifically from a smoke experiment (extracting 10 minutes of data from 4 days) and calculated the average signal. A few prominent ions had to be removed because of their constant production within the instrument (mostly related to water clusters) unrelated to the samples. The top 150 or so compounds by this average abundance were included in a select list of ions to analyze out of the original 2000+ (Main important mzs for analysis v3.xlsx). Several further signals with slightly lower abundance were included because either I had some confidence in their identification or their formula were sufficiently simple as to be expected in the data. For example, certain hydrocarbons (i.e. compounds only containing carbon and hydrogen atoms) did not make the 150 compound cut-off, but were clearly important and had unique m/z values.

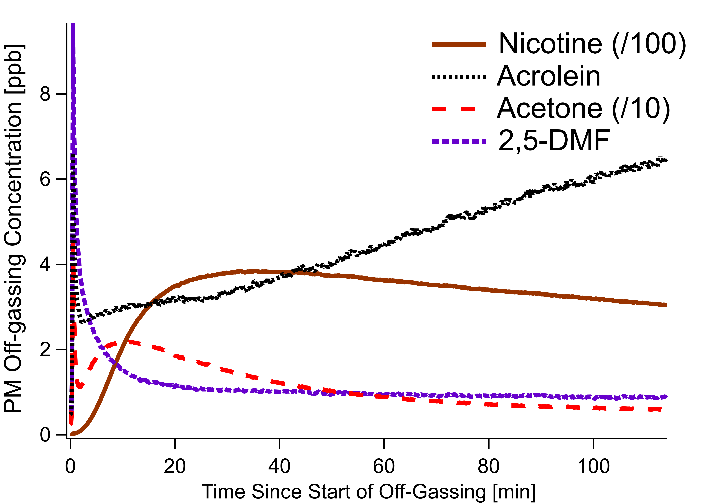
Further care had to be taken for larger compounds with higher molecular weights. These compounds are generally found in smaller concentrations in the atmosphere and in emissions, but they can play an important role in physical/chemical processing, reactivity, and toxicity. Simple outlier analysis was done for these. In addition to searching for chemically simple formulas, I used the local median factor, which is similar to the median absolute deviation. For each m/z, I compared its abundance to the median abundance for its 15 nearest neighbors (7 below, 7 above, and itself).

After these particular extractions, I ended with a list of 369 masses, each with an assigned chemical formula and approximately half with putative chemical identifications. Many others had fairly simple formulas that could be any number of isomers, so putative identifications were not assigned.

*Further Data Transformations*

I resampled the original data set from 1 second to 10 second data, then extracted only the 369 masses in the list I had compiled, using code that essentially acted as a query (a la SQL). However, the data were still reported as signal (ions per second). Therefore, I needed to determine sensitivity (i.e. a factor that converts these signals to concentrations in ppb) for each of these compounds. The sensitivity was calculated from the calibrant gases, using a method devised by Kanako Sekimoto and Jordan Krechmer (Aerodyne) that will be outlined in my paper in further detail. These sensitivities were applied to the time series.

Since the Vocus PTR-TOF-MS runs continuously, but experimentation conditions were not always similarly continuous, I wrote code (below left) that took in some user input (start time and end time) to extract time series slices for further processing (e.g. pulling a 10 minute long period of data). These slices were directly graphed as time series (below right) with a heuristic for extracting representative concentrations, depending on the experiment type: (a) selecting the 95th percentile highest concentration (secondhand smoke or particulate-matter off-gassing) or (b) removing the first two minutes of sample and then selecting the maximum data (lung lining fluid off-gassing).



*Output of Cleaned Data - Summary*

Since starting with multiple 1.6 GB sized data sets containing 2000 ions with ion/second data on a 1 Hz frequency, after a number of transformations, I arrived at the most important 369 ions with parts per billion (ppb) concentration measurements resampled to one data point every ten seconds. In addition to reducing the file size, doing these transformations simplified the data by removing extraneous information. For each experimental trial, I also calculated representative concentrations, which allow us to compare the behavior of these compounds across different experimental conditions (i.e. A/B testing). We also calculated more metrics based on compound identifications, specifically volatility and Henry’s Law Constant, which allowed us to find trends in compound behavior depending on these properties. More details on this procedure and other analyses carried out for the paper will be in the eventual manuscript.