Geotracker GAMA Chemical Analyzer

Sidd Singal, Subhash Gubba

May 7, 2015

Abstract

Hexavalent Chromium (CrVI) is a toxic chemical which can be found in groundwater across the United States. Using a public dataset that contains information about chemical readings in California groundwater wells, we created a tool to explore visualizations and analysis that could lead to important findings in the correlations between concentrations of chemicals in California groundwater. We also explore a case study which was performed using our tool, focusing on CrVI and its relationships to Chromium in Yolo County, California.

Hexavalent Chromium ($CrVI^+$ or CrVI) is a carcinogen and mutagen found in water systems in areas where industrial activity is prominent [1]. It is estimated that half a million workers in the United States are exposed to CrVI[2], yet the question still remains how we can effectively deal with its contamination in groundwater wells. The Environmental Protection Agency set a maximum contaminant level (MCL) of 100 parts per billion (ppb) for CrVI in drinking water[1]. However, California has a history of CrVI contaminants in both air and groundwater, and California set its own MCL of 10 ppb for CrVI in drinking water. Despite having a lower MCL than the one mandated by the EPA, California would like to meet an MCL of .02 ppb[3]. A particular goal of Stanford University's Department of Environmental Earth System Science (EESSD) is to track and understand the propagation of contaminants. Using this information, the EESSD wants to find ways to improve regulation and treatment of groundwater[4]. With assistance from Annika Alexander-Ozinskas from the EESSD, we built a tool (shown in Figure 1) that helps researchers find correlations among any chemicals that are measured in California's wells. Using our tool, the government of California can gain new insights on how to control some chemical X's contamination if they discover that X and another chemical Y are associated with each other, or if X concentrations are more prominent during certain time periods/events or locations.

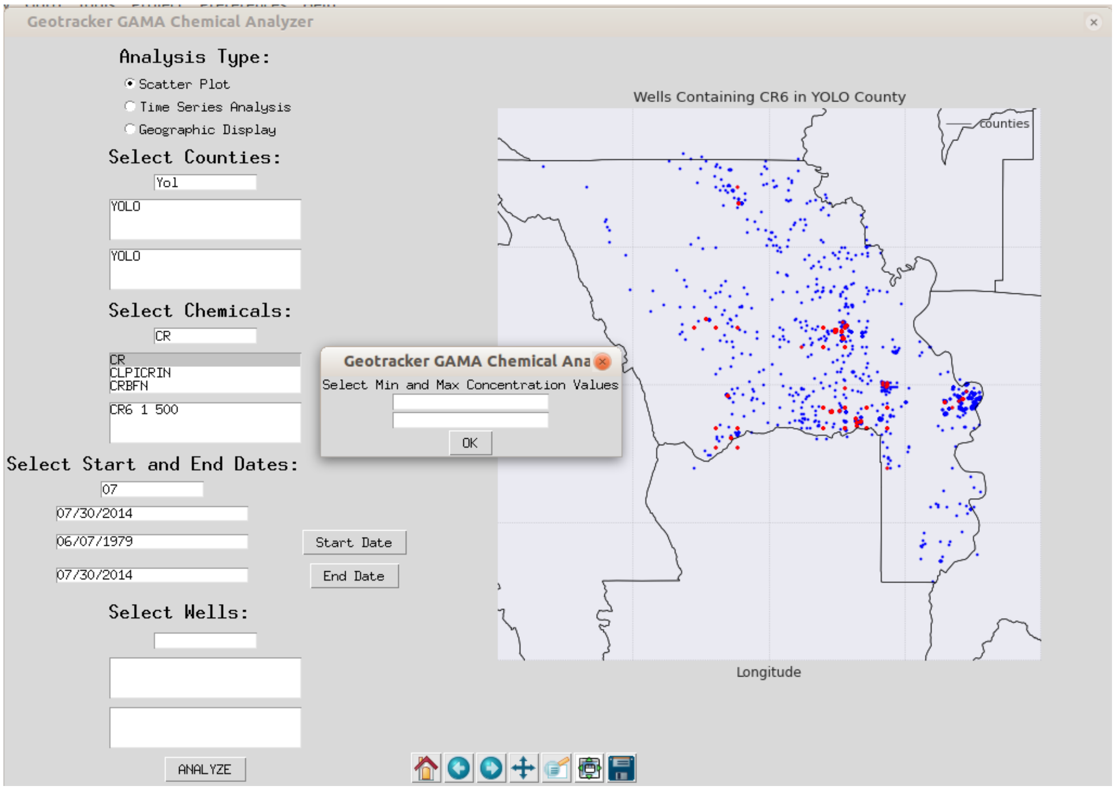

Figure 1 *This is an example of what our interface looks like. The control pane shows up on the left side, while graphs show up on the right. The bottom is reserved for any special messages and calculations.*

The dataset we are working with comes from the Geotracker GAMA web-based tool, provided by the State of California, which contains chemical content readings from groundwater wells located throughout California. The data includes information regarding date, chemical name, geographic location, and concentration[5], allowing us to combine and analyze the information in a number of contexts. We achieve this through a tool we built called *Geotracker GAMA Chemical Analyzer* (GGCA), developed with the Python programming language. Data is stored in a MySQL database to optimize the information retrieval process and storage. The MySQL database also allows us to store the data on the hard drive as opposed to in memory. The dataset is ~6.3 GB, which is generally too large for many computer RAMs to handle.

Figure 1 depicts GGCA upon launch. The left pane of the tool contains a set of controls for the user to adjust so results can be displayed and filtered however they are prefered. First, the user chooses the desired graph. The graphs include:

- Scatter plots with regression: plot concentrations of a chemical in question versus the concentrations of another chemical in question. The scatter plot includes a regression line and numerical information regarding the regression line.
- Time series analysis: for each chemical in question, plot the concentration of the chemical from a specific well over time.
- Geographical visualization: for the chemical in question, show its concentrations across the given county/counties based on location.

The control pane also includes the option of filtering data by date, so the user can select a beginning and ending date. From there, the user selects which counties to gather data from and which chemicals to analyze. For the time series analysis, the user is not limited to the number of chemicals selected, but the user will only be able to select two chemicals for the scatter plots and multiple chemicals for the geographical visualization. There is also a panel where the user can select which wells to include data from.

To demonstrate the software, we present a case study discovering and examining potential relationships between Chromium (Cr) and CrVI in Yolo county. It is possible that concentrations of Cr are directly linked to the concentrations of CrVI. If there is a link, then the government of California can introduce policies to regulate Cr concentrations with the hopes of controlling CrVI concentrations as a side effect. One method of checking for correlation between the concentrations of the two chemicals is to generate a scatter plot of the concentrations, as shown in Figure 2. Upon initial inspection of the plot, there is not an evident correlation between the CrVI and Cr concentrations. The Least Squares Regression Line (LSRL) is almost horizontal, indicating that we have minimal predictive power of CrVI concentrations given a Cr Concentration. However, a substantial majority of the Cr readings are below 100 μg/L, even though the plot shows a range up to ~650 μg/L. Using only data with Cr concentrations below 100 μg/L, we see a positive correlation between the two concentrations. Using GGCA, we limit the range of Cr concentrations to be between 0-100 μg/L. This new scatter plot is shown in Figure 3. The general trend is that a higher Cr concentration tends to result in a higher CrVI concentration. The new LSRL has an equation of CrVI = 2.85 + 0.69(Cr). This means that for every additional μg/L of CrVI, Cr concentrations will rise 0.69 μg/L. GGCA also tests the strength of this relationship by displaying the Spearman's correlation for any scatter plot. In this scenario, GGCA calculated Spearman's correlation to be 0.86, which is evidence that there is a strong, positive correlation between Cr and CrVI concentrations. Looking at these results, the EESSD or California may try to push for policies that limit Cr contamination in the groundwater. The thresholding we performed also gave us insight that our tool should have a feature which allows the user to select minimum and maximum concentrations for the chemicals, in case they want to work with specific portions of the data.

Though the scatter plot information is useful, our dataset gives us the power to produce other visualizations that involve date information. The dates assist us in doing a time series analysis of any of the chemicals we wish to look at. Using the GGCA, we view changes in concentrations of Cr and CrVI in the MW-1 well over time in Figure 4. We isolate this well as it contains a high amount of data spanning a 10 year time window and could serve as a good representation of the surrounding areas. There is a remarkable spike in the concentrations of both chemicals in 2008. In addition, There are also recurrent spikes in CrVI concentrations throughout the years. Researchers may use this time series analysis to investigate major events that occurred in 2008 (and other years where spikes have occurred). For instance, EESSD might find that Yolo county had major construction happening in 2008, and they may decide to push for policies that limit pollution from construction sites.

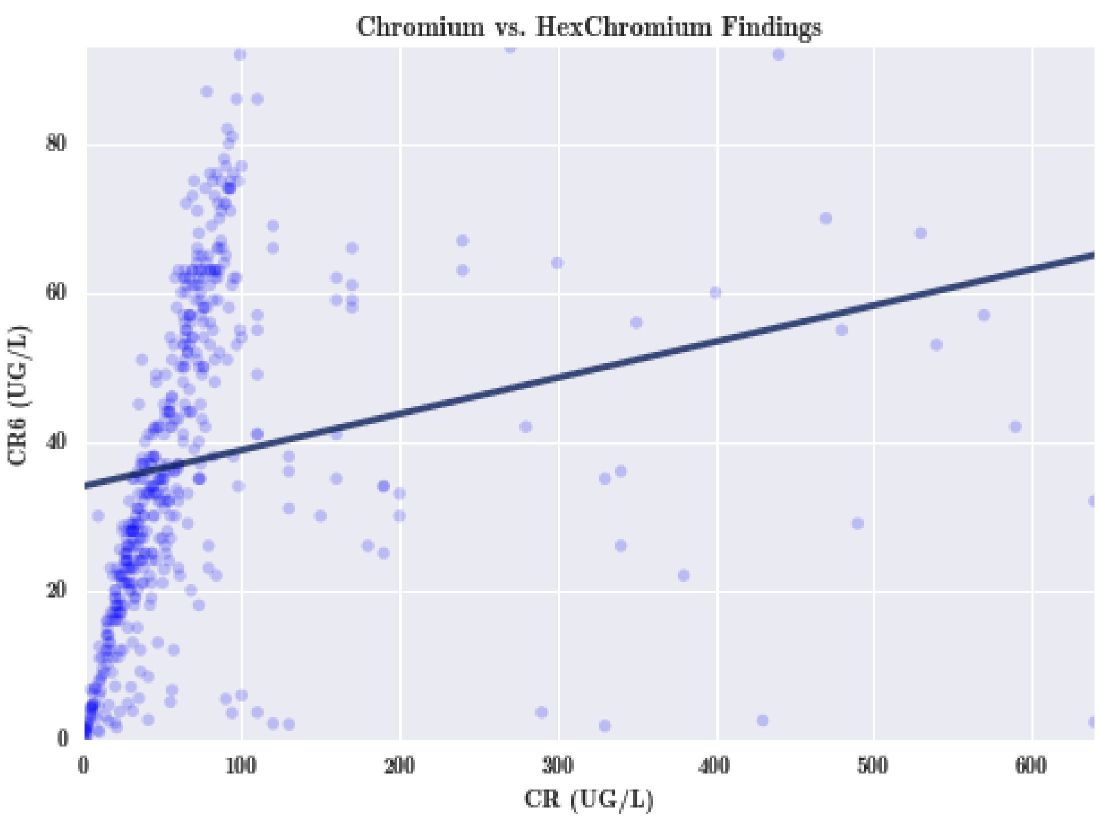

Figure 2 *Scatter plot when hexavalent chromium concentrations from Yolo county are plotted against their corresponding chromium concentrations.*

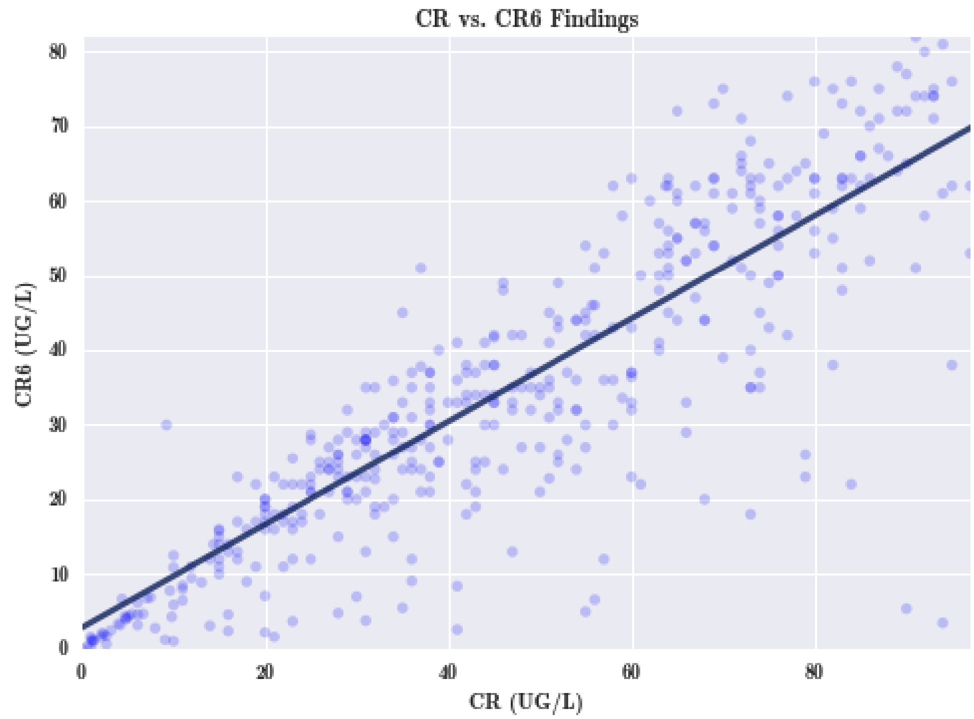

Figure 3 *Scatter plot with the same data as Figure 2, except data with Cr concentrations greater than 100 µg/L are not considered when fitting the LSRL.*

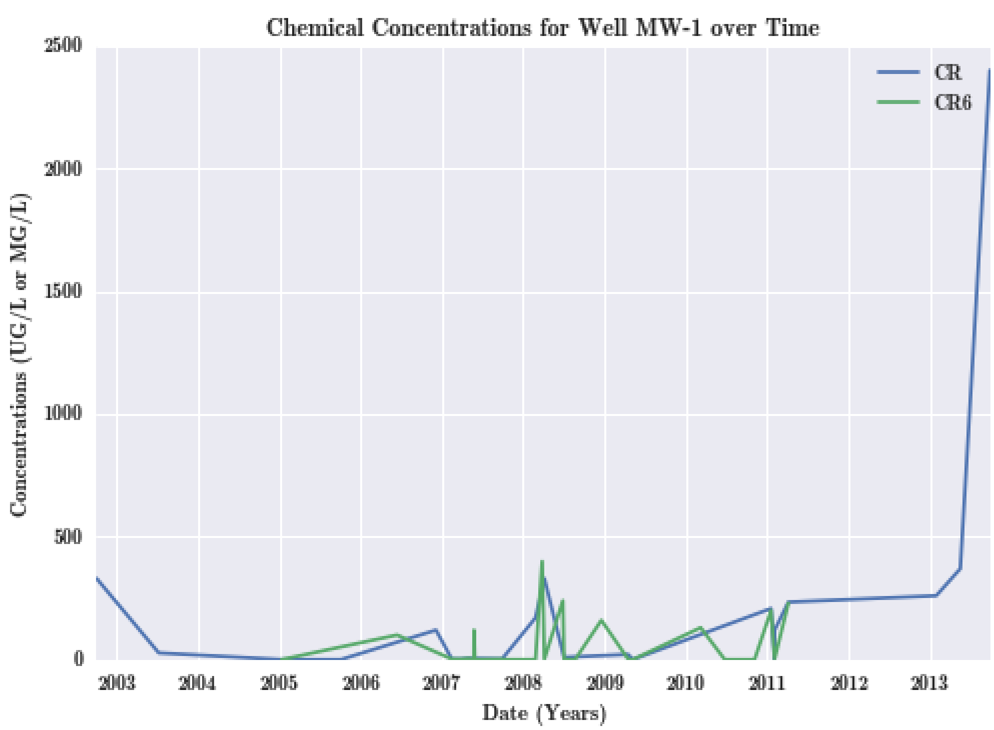

Figure 4 *Time series of Cr and CrVI in the MW-1 well in Yolo County.*

California also needs to determine where CrVI contamination is actually a problem. It may not be worthwhile to investigate areas where the contamination is either minimal or nonexistent. GGCA is able to locate exactly where any CrVI contamination is measured. Figures 5 and 6 show how maps of the data can be useful: for showing clusters of wells (blue being all wells, and red being wells where only CrVI was found), and how those concentrations vary by location (Figure 6, white indicating lower concentrations and darkening as they increase). Based on these figures, EESSD may decide to target their research to the more southeastern portion of Yolo county and determine why that region tends to have higher CrVI concentrations.

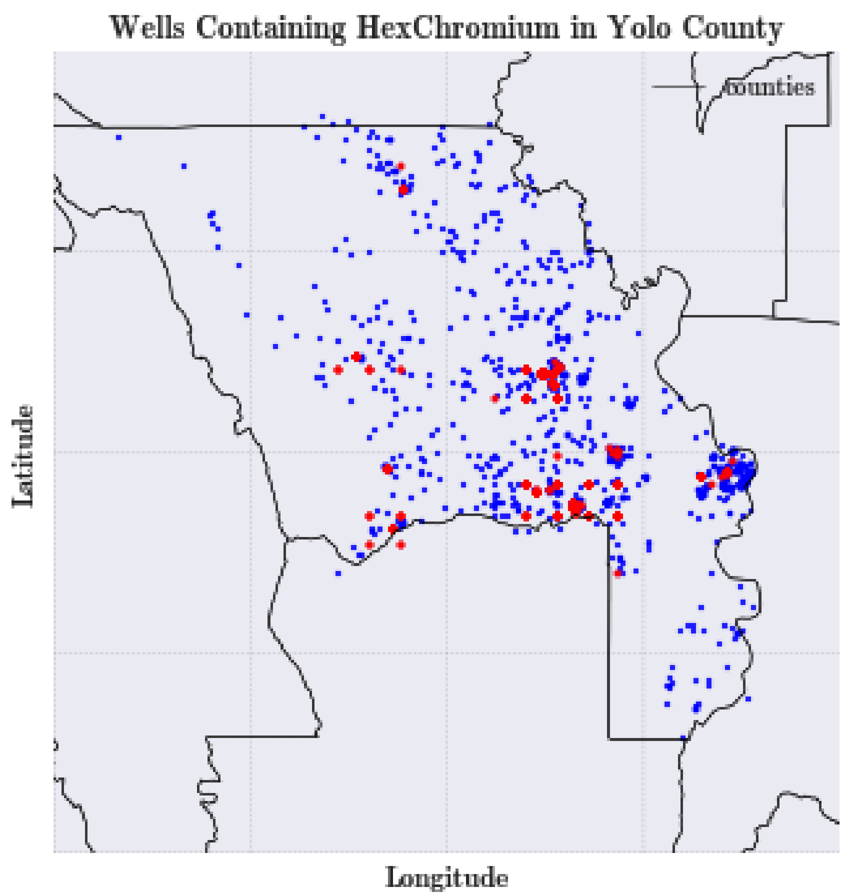

Figure 5 *Graph of all wells in Yolo County (blue) and wells where CrVI can be measured (red). The lines on the graph divide the California counties.*

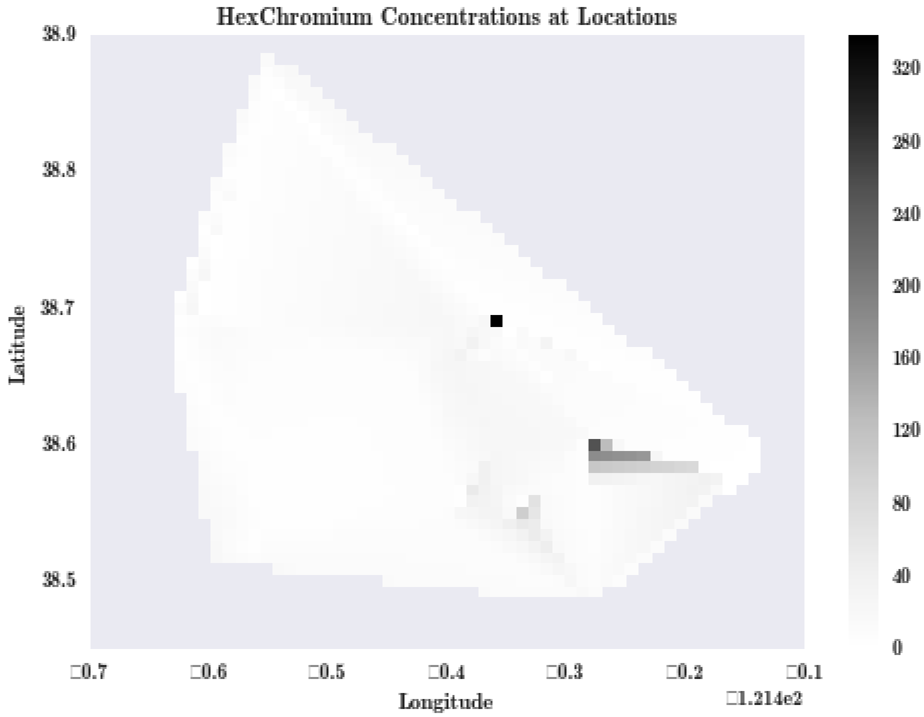

Figure 6 *A heat map indicating the concentrations of CrVI spatially. Darker regions on the map represent where higher concentration levels were found in wells, while white is lower.*

In conclusion, GGCA can help find insights into the changing state of contaminants in California's groundwater system. With the power of our tool, we hope the EESSD will be able to better pursue their mission of curating knowledge around the subject of groundwater pollution. We were able to harness California groundwater data to create statistical analysis which we integrate into a tool for use by the EESSD. Some further directions we could take this project are to tap into the power of web resources to host the tool publicly, apply our system to other groundwater data from around the country, and perhaps integrate with Google Maps or ARC GIS (Geographic Information Systems tool) to make a tool which would have both statistical and advanced geographic prowess, as a type of augmented GGCA software.

References

[1] Basic Information about Chromium in Drinking Water. (2013, December 13). Retrieved April 12, 2015, from http://water.epa.gov/drink/contaminants/basicinformation/chromium.cfm

[2] Hexavalent Chromium. (n.d.). Retrieved April 12, 2015, from https://www.osha.gov/SLTC/hexavalentchromium/

[3] Office of Environmental Health Hazard Assessment,. (2011). *Hexavalent Chromium PHG*. Retrieved 12 April 2015, from http://www.oehha.org/water/phg/072911Cr6PHG.html

[4] Stanford Earth System Science. (2007). Retrieved February 2, 2015, from https://earth.stanford.edu/departments/eess/

[5] (2011). Geotracker GAMA Groundwater Information System. *California EPA State Water Resources Control Board*. Retrieved February 2, 2015, from http://geotracker.waterboards.ca.gov/gama/gamamap/public/default.asp