Code Data Integration 19-07-2021

Abishek Laxmanan Ravi Shankar

1. Hypothesis

The chemicals dissolved in the water in rivers are one of the major reasons for the allergies and for the dermal problems.

2. Aim

To compare and contrast the datasets with programming, thereby obtaining significant values and evidences in order either prove or disprove the hypothesis.

3. Data Provided

5 '.csv' files are provided from the CTD database.
(http://ctdbase.org/about/publications/#citing) [6]

The names of these files are as follows:

- **CTD_chemicals.csv** → Comprises of chemical names and the corresponding MESH IDs.
- **CTD_disease.csv** → Comprises of disease name and the corresponding MESH IDs
- **CTD_chemicals_diseases.csv** → Comprises of both chemical names and the corresponding disease names and the MESH IDS
- **empodat_surface-water_filtered-geo-concentration.csv** → Comprises of Norman_SusDat_IDs and the concentration of the corresponding chemicals dissolved in water.
- **susdat_chemical_information.csv** → Comprises of chemical names to the corresponding Norman_SusDat_IDs. It also possess the different chemicals toxicity levels and their exposure.

4. Materials

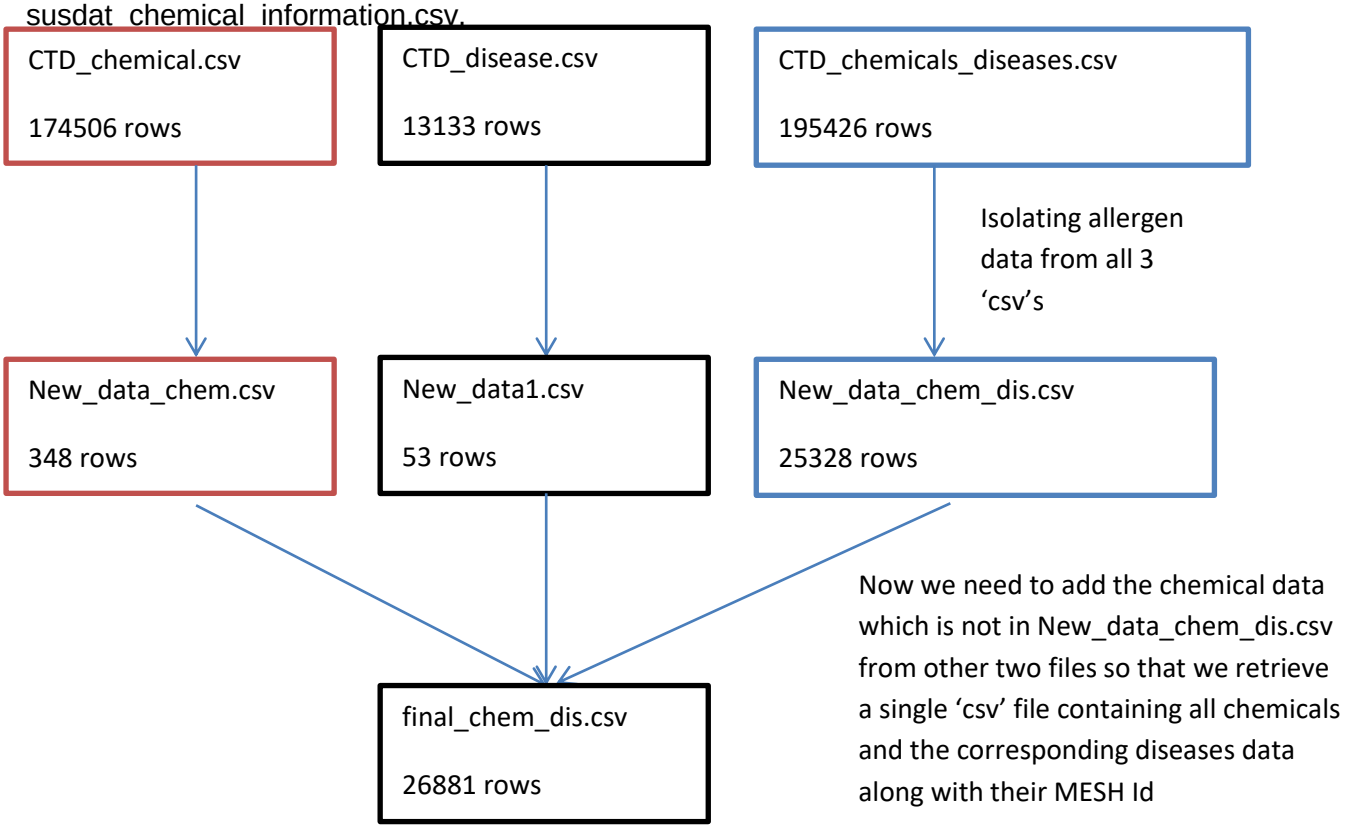
- Jupyter Notebook → Python3 ; to operate with the '.csv' files [4]
- Libre Office Calc → To view the '.csv' files [5]

5. Methods

5.1 Overall Workflow

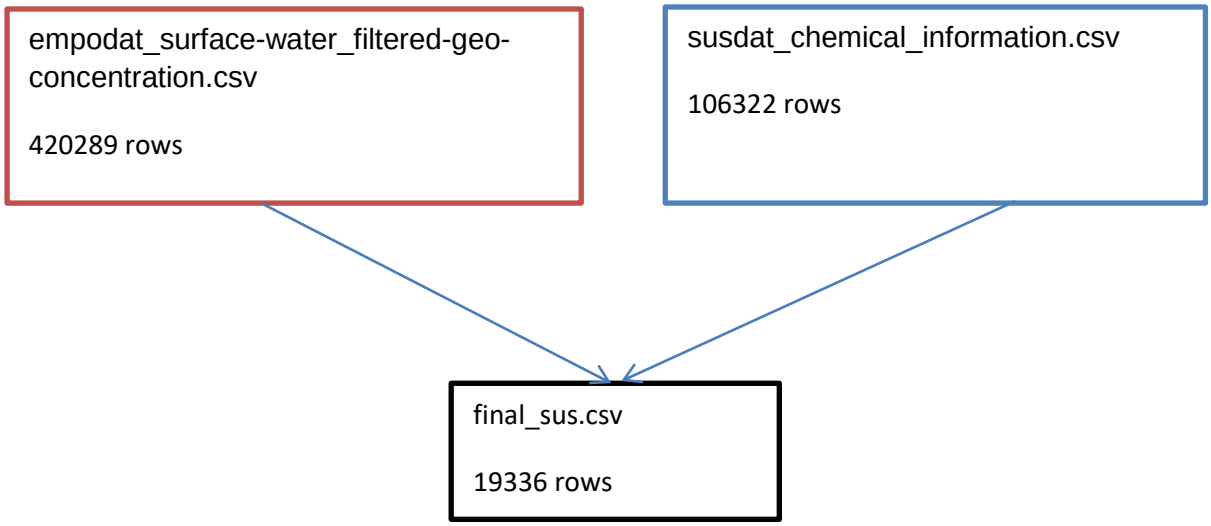
In order to perform data analysis for the allergen data, the corresponding allergy data alone must be initially isolated to perform faster comparison.

On the other hand the “empodat_surface-water_filtered-geo-concentration.csv” comprises of dissolved concentration which is sorted and isolated and is merged with the susdat_chemical_information.csv.



(Fig 1: Combining CTD datasets into a single final_chem_dis.csv)

Now these 26881 rows contain only allergy and related diseases data. This has to be compared with the dissolved concentration data.



(Fig 1: Combining SusDat into a single final_sus.csv)

Now the files to be compared is ready as.

- final_chem_dis.csv possess the chemicals and the corresponding allergy related diseases,
- final_sus.csv possess the chemicals with water dissolved concentrations.

These two files are compared based on the chemical names and the common chemicals which are dissolved in water are returned. They are as follows:

['endosulfan sulfate', 'cyprodinil', 'isoproturon', '2-nitrophenol', 'pyrene', 'iomeprol', 'bentazone', 'cyproconazole', 'chlorpyrifos-methyl', 'pendimethalin', 'diethyl phthalate', '4-chlorophenol', 'tebuconazole', 'propylparaben', 'propachlor', 'tributyl phosphate', 'methyl tert-butyl ether', 'aniline', 'abamectin', 'prochloraz', 'phenanthrene', 'mecoprop', 'diflufenican', 'benz(a)anthracene', 'terbutylazine', 'carbendazim', 'quinoxyfen', 'propiconazole', 'oxadiazon', 'bisphenol a', 'methylparaben', 'metolachlor', 'glyphosate', 'metaldehyde', 'azoxystrobin', 'diisononyl phthalate', 'naphthalene', 'chloramine-t', 'fluoranthene', 'imidacloprid']

This in return returned 45 rows from 'final_sus.csv'.

The table containing chemical name and the corresponding dissolved water concentration is as follows.

Table 1: The chemicals which causes allergy and their dissolved concentration in river water

	Norman_SusDat_ID	**Name**	**Dissolved Concetnration (µg/l)**
106	NS00000117	cyproconazole	0.081
107	NS00000118	cyprodinil	0.87
201	NS00000220	propiconazole	4.7
209	NS00000228	iomeprol	0.36
217	NS00000236	bentazone	30
229	NS00000248	metolachlor	52
239	NS00000258	terbutylazine	30.4
241	NS00000260	isoproturon	3.19
258	NS00000277	mecoprop	0.33
261	NS00000280	tebuconazole	0.99
303	NS00000323	propachlor	0.137
341	NS00000361	imidacloprid	1
425	NS00000446	chlorpyrifos-methyl	1.6
507	NS00000529	pendimethalin	0.077
1778	NS00001849	methylparaben	0.22

2039	NS00002126	propylparaben	0.23
2793	NS00002920	diisononyl phthalate	81.26675
6025	NS00006349	glyphosate	17
6639	NS00006999	metaldehyde	1.5
7876	NS00008307	methyl tert-butyl ether	3.8
7945	NS00008377	naphthalene	0.424
8010	NS00008448	oxadiazon	0.3
8381	NS00008837	diflufenican	0.28
8408	NS00008865	bisphenol a	21
8994	NS00009484	phenanthrene	126.1
9278	NS00009778	azoxystrobin	0.49
9387	NS00009893	diethyl phthalate	1.75
9482	NS00009999	endosulfan sulfate	0.029
9724	NS00010255	tributyl phosphate	18.42
9734	NS00010265	carbendazim	6.6
9878	NS00010414	4-chlorophenol	0.11
10042	NS00010580	quinoxyfen	0.043
10067	NS00010607	2-nitrophenol	0.06
10113	NS00010656	aniline	0.13
10157	NS00010700	fluoranthene	5.7
10158	NS00010701	pyrene	5.1
10178	NS00010721	benz(a)anthracene	33.8
10338	NS00010885	prochloraz	0.68
13640	NS00014275	abamectin	0.05
64381	NS00066780	chloramine-t	0.2

Table 2: The chemicals which causes allergies 0 concentration in the rivers water

	Norman_SusDat_ID	Name	Dissolved Concetnration (µg/l)
95332	NS00100095	abamectin	0
94955	NS00098640	prochloraz	0
95445	NS00100410	abamectin	0
96303	NS00101930	tributyl phosphate	0
100913	NS00108197	abamectin	0

6. Analysis and Observation

So now I compare the chemical names to their concentrations. A research paper named "Maximum Contaminant level" [3] states that the maximum concentration of Nitrile compounds in water can be (1 mg/L) [2] and the arsenic in composition can be upto 0.1 (mg/L) [1] .

The concentration obtained from the 'file.csv' do not possess any values in phenanthrene (126.1 µg/l) and diisononyl phthalate (81.26675 µg/l) which is less than 1 mg/l therefore marking the river water to be safe, thereby disproving the hypothesis.

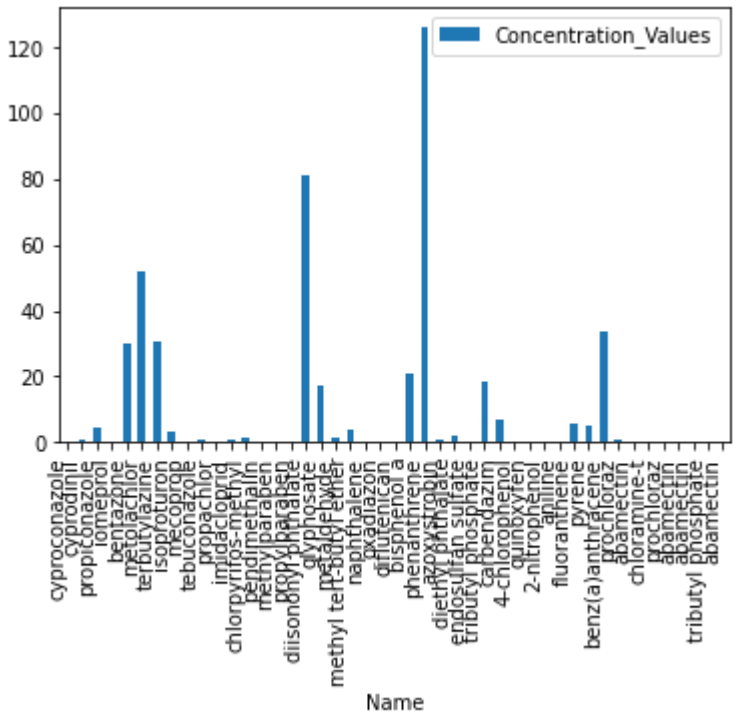


Fig 3: Chemical name vs Concentration values (µg/l) found in river water

7. Conclusion

Since the dissolved concentration of chemicals which causes allergies are in way too less dissolved concentration in the river water and it is much lower than the safety standards, it is conclusive that the chemicals in river water is NOT a predominant reason for the allergies, hypersensitivity and other dermal problems.

8. Reference

1. Santhini Ramasamy, Janice S. Lee, in Handbook of Arsenic Toxicology, 2015 → Arsenic Risk Assessment

2. J.M. Curl, ... W.D. Bellamy, in Comprehensive Water Quality and Purification, 2014 → Assuring Purity of Drinking Water

3. Maximum Contaminant Level; From → Encyclopedia of Toxicology (2014) https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/maximum-contaminant-level

4. Python3 in Jupyter Notebook → Fernando Perez, Brian E Granger, and John D Hunter. Python: an ecosystem for scientific computing. *Computing in Science \\& Engineering*, 13(2):13–21, 2011. https://jupyterbook.org/content/citations.html#id6

5. Libre Office Calc → Foundation, T. D. (2020). *LibreOffice Calc*. Retrieved from https://www.libreoffice.org/discover/calc/

6. CTD database → Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, Mattingly CJ The Comparative Toxicogenomics Database: update 2021. Nucleic Acids Res. 2020 Oct 17. http://ctdbase.org/about/publications/#citing

9. Annexure

The files included are as follows.

1) **Code data integration.ipynb** → Jupyter notebook containing code to perform string manipulation
2) **final_sus.csv** → Contains dissolved concentration data from empodat_surface-water_filtered-geo-concentration.csv database added to the 'susdat_chemical_information.csv' database
3) **final_chem_dis.csv** → Contains the combined MESH IDs, disease and the chemical data from CTD_chemicals.csv, CTD_disease.csv, CTD_chemicals_diseases.csv
4) **file.csv** → Contains the 45 chemicals which causes allergy along with their dissolved water concentration.
5) **Code Data Integration.pdf** → Explaining hypothesis and methods for the task.
6) **new_data.csv** → Scaled down CTD_disease.csv file to rows containing allergies
7) **new_data_chem.csv** → Scaled down CTD_chemicals.csv file to rows containing allergens related chemicals
8) **new_data_chem_dis.csv** → Scaled down CTD_chemicals_diseases.csv file to rows containing allergies and their corresponding chemicals

9) **water_data.csv** → Comprises of only compounds containing concentration greater than 0 from the empodat_surface-water_filtered-geo-concentration.csv