**Dataset**

This dataset[[1]](#footnote-1) is collected by European Monitoring Centre for Drugs and Drug Addiction, and focuses on drug-related indicators in municipal wastewater from selected European cities. The dataset consists of 2 files, one with location information and the other with indicators of the relevant components in the wastewater. After data alignment and data fusion, the two files are merged into one table, and the data examples are as follows:

Table 1 Example of data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| year | metabolite | country | city | Daily mean | Weekday mean | Weekend mean |
| 2021 | amphetamine | AT | Graz | 38.57 | 40.03 | 37.48 |
| 2021 | amphetamine | AT | Hall-Wattens | 6.97 | 7.74 | 6.39 |

The quality and information of the data are as follows:

Table 2 The quality and information of the data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Year / country / city | Daily mean | Weekday mean | Weekend mean |
| count | 1650.000000 | 884.00 | 883.00 | 884.00 |
| mean | / | 260.81 | 233.47 | 281.76 |
| std | / | 516.32 | 465.50 | 559.89 |
| min | / | 0.00 | 0.00 | 0.00 |
| max | / | 3680.13 | 3283.00 | 4279.81 |

First of all, we can see that the main indicators are the daily mean, weekday mean and weekend mean of the amount of different metabolites in the water stream of different places.

Two main dimensions can be sorted out from the data: geographical location and time dimension. For time dimension, there are several important times or timelines: 1) last year; 2) all years examined. Last year's data can help us understand the latest situation, and all the detected years can help us understand the overall situation of each region. From a geographical point of view, the data table involves a total of 100+ locations of 30 countries. Therefore, it is necessary to organize the data of various places into national data, which means merging testing data from different places within the same country to the national level.

**Data preprocessing**

As mentioned before, we need to aggregate the levels of different cities in a country to represent the level of drugs in this country, which not only makes the dataset more readable also alleviate the data missing problem (can be seen from table 2). There are some computed variables used in this dataset that need to be clarified:

It is a common practice to use the average value of different cities in a country to represent the drug level in this country. Here, we use to indicate the country, means that the cities belong to country , for example, . And is the detected ingredient, in this case, . In that way, the average of content of metabolite in country can be presented as:

Where is the detected figure of in city , and is a characteristic function, the definition is:

Due to differences in populations across countries, a comparison of detected metabolites alone is not sufficient to conclude that a country has a severe or better drug problem. Assuming the severity of drug abuse is the same in both places, the use and emissions of related substances must be higher in places with higher populations. So here, we introduce the concept of per capita emissions :

Where is the number of populations of the city .

After that, we can get each metabolite’s international average content of daily, weekday and weekend, also per capita emissions of daily, weekday and weekend. Here, we can also obtain the international level of all metabolites through the average number of all metabolites:

Where indicates the number of valid component types:

Similarly, per capita emissions of all metabolites can be defined:

Through figure 1, we can see the difference among the mean daily, weekday and weekend indicators, where the abscissa is the time axis, and the ordinate is the annual average of detection value of various metabolites. We can see that there is not much difference between those 3 mean values, only slight gap in case of cocaine and methamphetamine. So in this report, I am going to use mean daily value instead of mean weekday value and mean weekend value, to represent the level of each country.

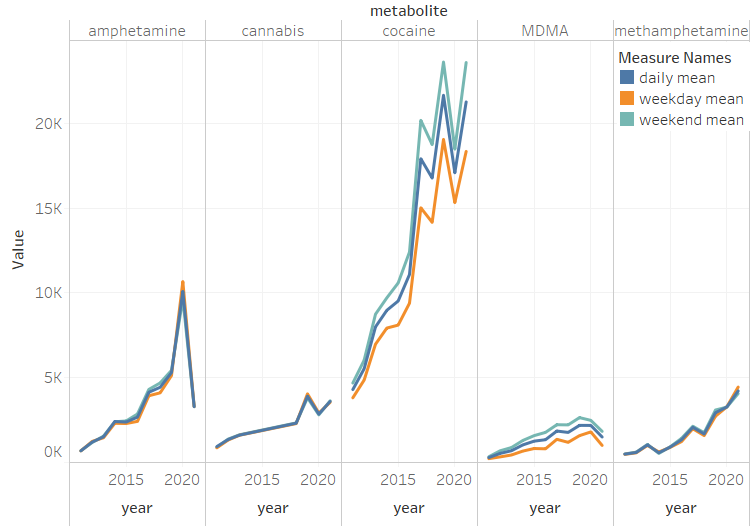


Figure 1 Annual average of detection metabolite value of various metabolites

**Data visualization**

On the one hand, I want to explore which area has the most drug abuse based on the data collected. In figure 2, the left side one is annual average density distribution map of . It can be seen from this that the degree of drug abuse in areas around the Mediterranean Sea is more serious than in other places. And right one is a more detailed country distribution map, the depth and shade of the color indicates the size of .

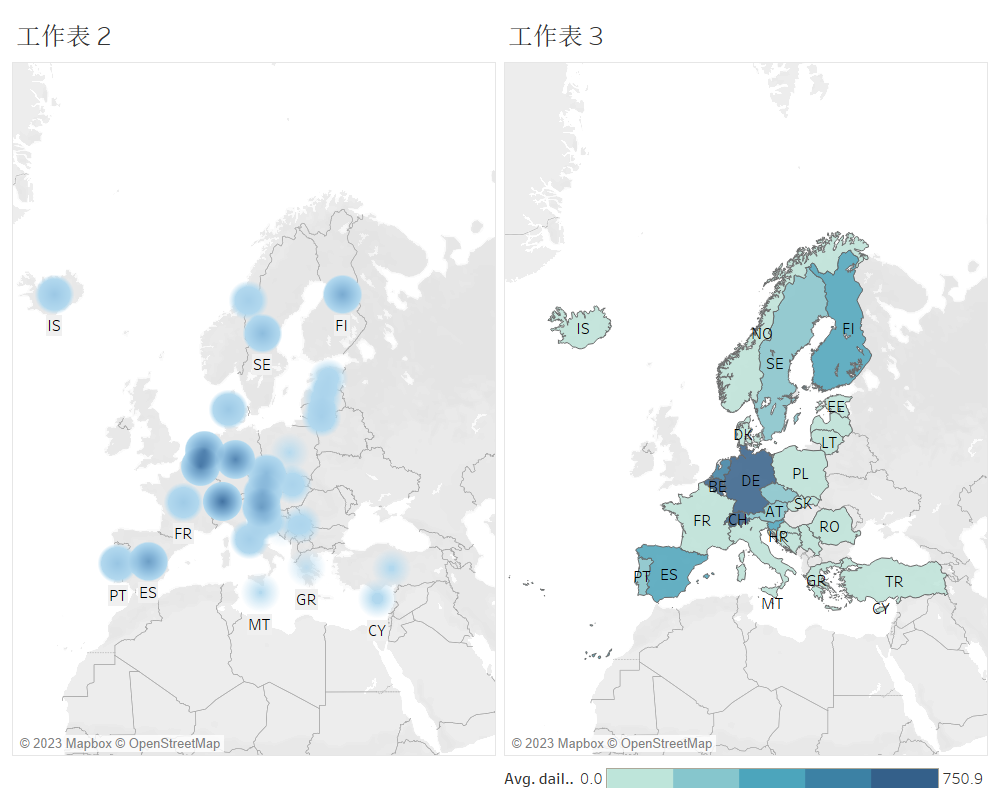


Figure 2 annual average distribution map of

We can also compare the drug levels of various countries in more detail through the bar chart Figure 3. From there, we can tell that drug problems in Switzerland (CH), Germany (DE), Belgium (BE), Netherlands (NL), Spain (ES) are relatively serious, with Switzerland being the worst country.

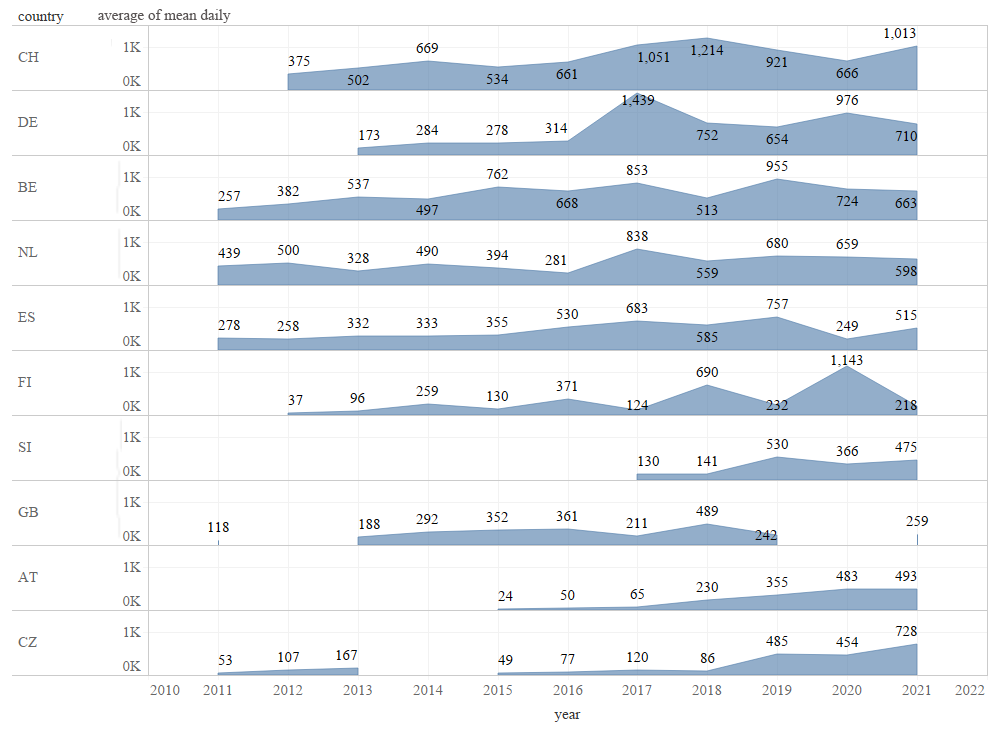


Figure 3 for each country over time (only show top10)

But it is easy to find that the above analysis does not take demographic factors into account. That is, countries with large populations and countries with small populations are considered to have similar levels of drugs if their indicators are similar. This obviously has some flaws. Therefore, taking into account the population factor, we again visualize, as figure 4-5. And we can discover that some countries are disguised by their small population in the previous analysis, such as Iceland (IS) and Slovenia (SI). So, we can consider that the lie factors of previous figure 2-3 are not equal to 1, according to Tufte’s principles.

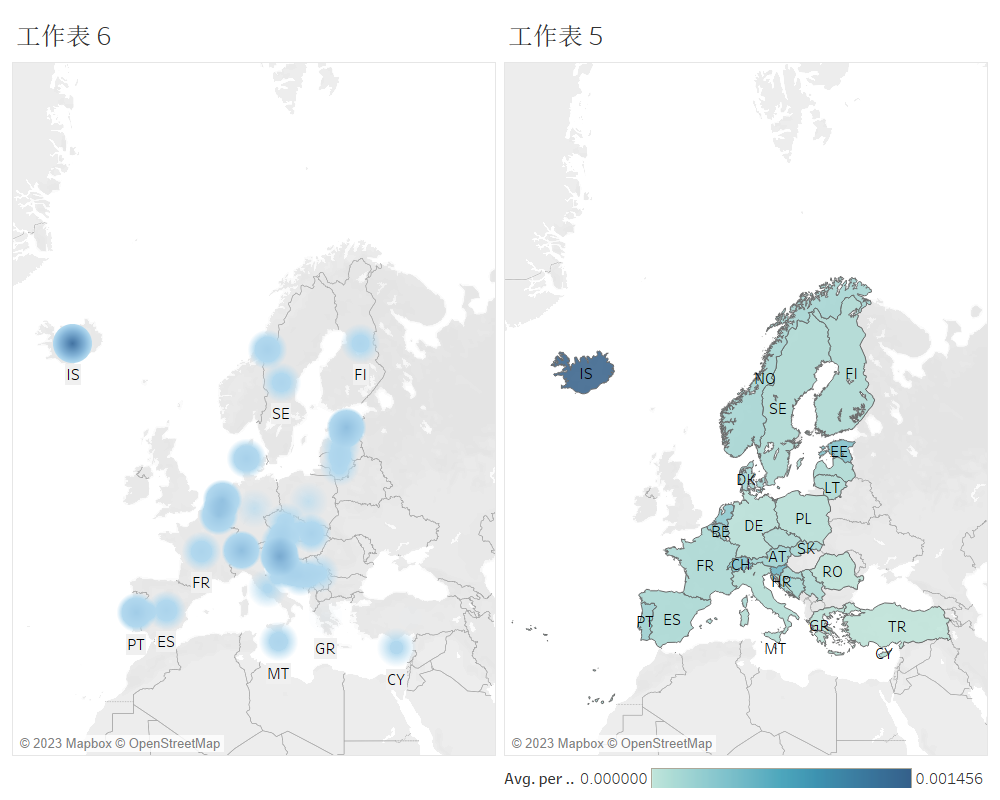


Figure 4 annual average distribution map of

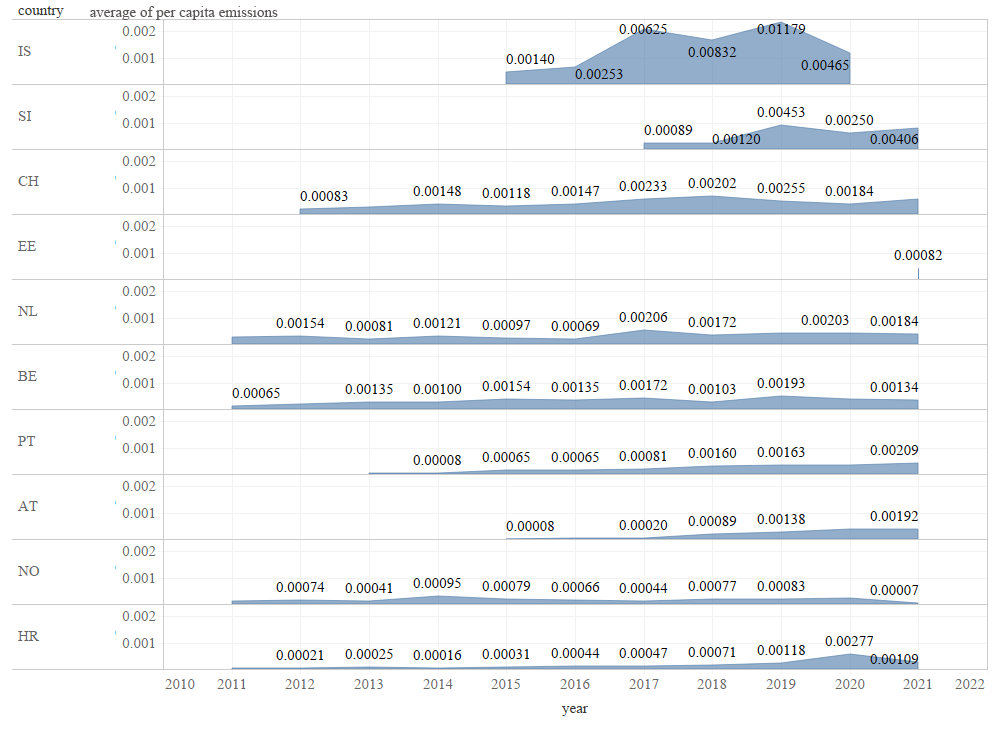


Figure 5 for each country over time (only show top10)

From the perspective of visualization, we can find that in Figure 2 and Figure 4, the depth and shade of the color directly reflects the severity, allowing people to intuitively understand the image in a pre-attentive way. Similarly, it can also be judged pre-attentively from Figure 3 and Figure 5 by coloring area acreage. Also, according to the Closure feature of Gestalt laws, each country can be easily seen as separate entities on the map. And Similarity of Gestalt laws also make sure that patches of the same or similar color will be considered to have similar levels of drug abuse in Figure 2 and 4.

1. https://data.europa.eu/data/datasets/drugs-in-municipal-wastewater-in-selected-european-cities?locale=en [↑](#footnote-ref-1)