Link between air pollution and fatal diseases

1152502_Jongho Park

1181873_Dulan Wijeratne

1164216_Marco Altieri

1171273_Aidan Prescott

Introduction

Air pollution is a major concern in society today. It is often associated with negative health implications. Therefore, we decided to take a further look at this and explore if there is any actual proof that there is a correlation between diseases that cause and the increase in the air pollution of Melbourne. To keep with the theme, we focused on the health of the Victorian residents. Which leads me to our research question, does the increase in air pollution cause more fatal diseases in the residents of Melbourne. Specifically we wanted to look at the following diseases: Asthma, Chronic obstructive pulmonary(COPD), Heart failure, Ischaemic heart disease and skin cancer. We took two diseases that relate to the pulmonary system, two diseases that relate to the heart and one relating to the skin because we wanted to see what type of disease was more likely to occur, as well as the rate each disease was happening and what specific chemical was causing each of these diseases.

Linking dataset

The first dataset looked at the number of deaths in each region due to specific diseases. The second data set looked at the amount of different chemical pollutants in the region's atmosphere. It is already known that pollution correlates with diseases but we thought if there is a pattern when it comes to death by the given diseases, perhaps indicating severity. The chemical data set gave names of suburbs such as Alphington, Dandenong and Geelong South, whereas the other data set gave names of the regions such as Eastern Metropolitan, Southern Metropolitan and Gippsland. To able to link the two datasets together we made the assumption that suburbs close to one another had approximately the same amount of chemical pollutants, therefore we took a suburb or a group of suburbs in that region and then used that to find the amount if chemicals in each region, hence assumed that the suburbs were representative of their region. For example Geelong South belonged to the Geelong region so we used the amount of chemicals in the atmosphere of Geelong South as a representation of the amount of chemical pollutants in the whole of Geelong.

Wrangling Method and Analysis

We found two data sets from two different organisations(Victoria-Government-Data-Centre, Victorian-Government-Department-of-Human-Services). The problem with these two data sets is that they have different temporal formats

- Air quality dataset gathers the number of deaths **over 5 years**
- Death dataset showed the amount of each chemical in the region **every hour.**

In the air quality dataset, there were columns:

- sample point id
- sp _name
- latitude and longitude
- sample date
- time basis id
- param id
- param name

The dataset had over 700,000 rows per file, the format slightly changed from year to year, we had to redefine some columns to maintain consistency and prepare for feature grouping. Feature selection was used to filter the relevant chemical and its measured values per region. The data was compressed into averages over 5-year intervals, grouped by suburb. We decided to group each suburb into its relevant region as some suburbs weren't present in the disease other dataset, we felt that it is better to do a 1-to-1 comparison between regions. Chemical values were normalized and units removed as they had similar magnitude.

Example_Table: NW-metro 2007-2011

param_id	value
API	0.021
CO	0.009
NO2	0.344
O3	0.604
PM10	0.667
PPM2.5	0.265
SO2	0.036

The disease dataset represented the death count over 5-year intervals. We decided to use "death rate standardised over 100,000" as it provided a readable range.

A combined csv data was made with columns

- Years
- Regions
- Chemical
- Chemical values
- Death rate
- Diseases

VISUALIZATION

Chemicals were set as x-values, six chemical ratio(SO2, NO2, O3, API, PPM2.5, PM10) were used and the five diseases(Ischemic heart disease, COPD, Heart failure, Asthma, skin cancer) were used as y-values. These were chosen by:

1. Researching leading correlation between pollution and diseases
2. Omitting human caused deaths, such as suicide and assault
3. Chronic diseases are likely correlated by pollution

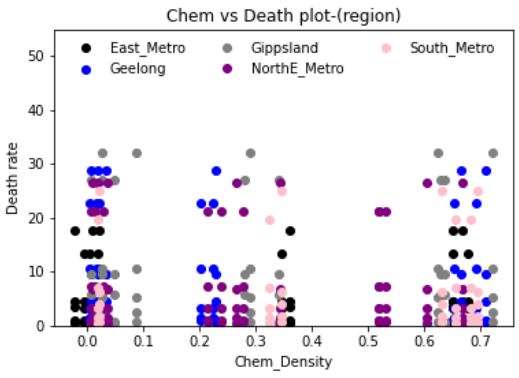
We wanted to see the death rate with chemical distribution by region. And the levels of scatter plots are shown and represent specific regions horizontally in Scatter_plot_A. Also in the disease plot(Scatter_plot_C), it shows a similar shape but a different thickness of every level that refers to a certain disease. Therefore the relation between the chemical ratio and the ratio of deaths of a certain disease can be assumed as they have similar features.

In Scatter_plot_A, we set black, blue, grey, purple, pink to East metro, Geelong, Gippsland, North East Metro, South Metro. Each region has similar chemical distribution in Scatter plot_B. In Scatter_plot_B, we set the colors black, blue, grey, purple, pink, yellow to PM10, O3, NO2, API, SO2, PPM2.5. The vertical layers representing each chemical are shown.

[Each chemical has a certain constant density depending on the region, not by time. In other words, each region exhibits a constant chemical density interval that is not affected by time. This important finding will be the basis of the purpose which is finding the relation between diseases and a constant ratio of certain chemical materials over years because the effect on the body by chemicals is accumulated and shows the consequence(diseases) after a long time.]

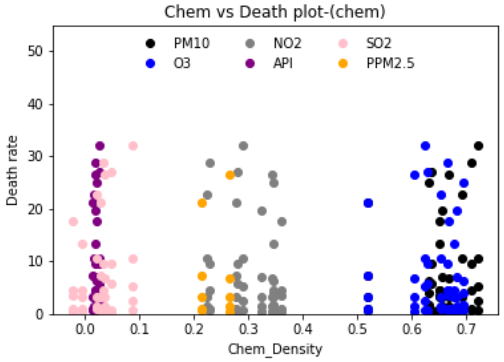
In Scatter_plot_C, we set black, blue, grey, purple, pink to Ischemic heart disease, COPD, Heart failure, Asthma, skin cancer. Black top level(Ischemic heart disease) is shown to be a major disease caused by pollutants as it has the highest death rate. and then COPD, Skin cancer etc.

SCATTER_PLOT_A



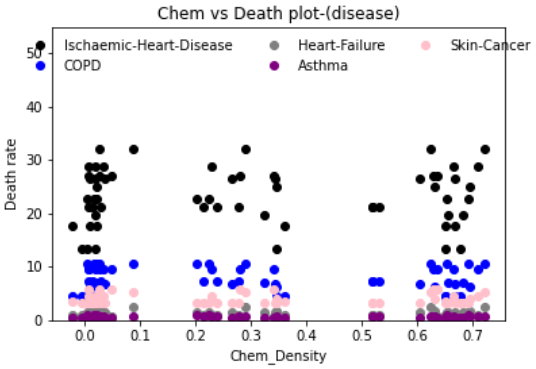
Plot of death rate against chemical density(normalized). A rather uniform distribution suggests that the chemical concentration is similar in each region. No apparent correlation can be observed between regional chemical distribution and death rate

SCATTER_PLOT_B



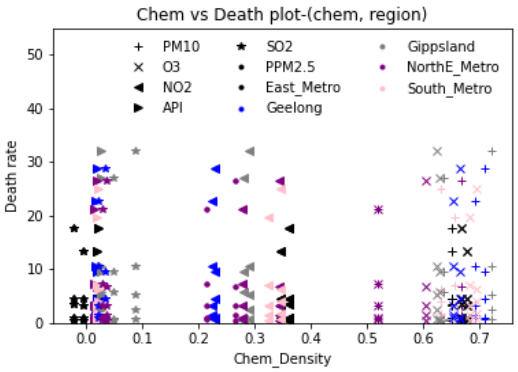
Plot of death rate against chemical density(normalized), with points identified by region chemical type. Plot suggests high toxicity of Sulphur Dioxide(SO2) and low toxicity of Ozone(O3) and PPM10. Sulphur Dioxide gas is known to be a toxic industrial byproduct and fertiliser reactant(forming SO2 product).

SCATTER_PLOT_C



Points identified by cause of death. Clear horizontal clusters are present, indicating cause of death where heart diseases are the primary cause of death. Asthma is by far the lowest cause of death.

FINAL_PLOT



Aggregating plot, shape indicates chemical while color indicates region. Gippsland being a farming region makes large use of fertilisers which contain sulphur. Conversely metropolitan areas do not have factories nor farm land hence have low SO2 concentration.

Conclusion and Significance of Research

It is shown that each region has a specific death ratio. For example, Gippsland has the highest death rate, Geelong is the second and then North East Metro > South Metro > East Metro.

We see the same order in the chemical SO2. The highest value is in Gippsland and the lowest value is East Metro. This means that we can assume the relation between SO2 and Death rate. Also as those death rates are in the Ischemic heart disease level. Specifically, SO2 and Ischemic heart disease are likely related. SO2 concentration is generally low as it reacts with Oxygen(O2) to produce Ozone (O3), the synthesis of O3 via SO2 is apparent due to the similarity in spread (Scatter plot_b). Higher concentration of SO2 suggests highly industrialized areas or high use of fertiliser, indeed Gippsland has high soil infertility therefore makes heavy use of fertiliser . Studies have shown an association between SO2 and ischemic heart disease(Qin et al., 2016)

PM10 also has the same region order and that means PM10 can be related with death rate as well. On the other hand, the rest of chemicals do not show the linear relation with death rate by region.

To prevent avoidable deaths caused by PM10 and SO2 we would recommend that an action should be taken to reduce the emission of these gases especially in Gippsland.

Limitations

In our research we made the assumption that the amount of chemicals in the atmosphere were approximately constant in all suburbs, in a particular region. Therefore, the final result may change if the amount of chemicals in each suburb were taken into account. Furthermore, by combining suburbs into regions, in our regional groups we had more representative suburbs (e.g. North-West Metro had over 6 suburbs but Gippsland only 4), this discrepancy for small regions groups causes more variance as the sample size is much smaller.

We also failed to take into account other factors that may cause these diseases like for example the smoking population in a region, which can also lead to diseases in the lungs as well as heart disease. Elderly people are more susceptible to a death given a disease, it might be useful to also look at the positive diagnosis of these diseases rather than simply the death rate.

Datasets

http://vhiss.reporting.dhhs.vic.gov.au/ReportParameter.aspx?ReportID=28&TopicID=1&SubtopicID=16

Victorian-health-information-surveillance-system-reports on avoidable mortality in regions of Victoria.

https://discover.data.vic.gov.au/dataset/epa-air-watch-all-sites-air-quality-hourly-averages-yearly/historical

Bibliography

- Qin,G.,Wu,M.,Wang,J.,Xu,Z.,Xia,J.,&Sang,N.(2016). Sulfur Dioxide Contributes to the Cardiac and Mitochondrial-Dysfunction in Rats.Toxicological-Sciences, 151(2),334–346.

https://doi.org/10.1093/toxsci/kfw048