Spatial interpolation of particulate matter across Greater Sydney

Declan Stockdale - 112145549

14 November 2021

Introduction

The 2019 - 2020 bushfires season that ravaged New South Wales (NSW) was unprecedented in both intensity and scale with approximately 18.2 million hectares being burnt (Fire and Rescue NSW, 2021) . This led to the generation of enormous long lasting smoke plums comprised on fine particles (particulate matter, PM) which blanketed large areas of NSW. The largest population center in NSW, Sydney with a population of over 5.3 million people, was engulfed by smoke for numerous days over the bush fire period at times recording PM10 (particulate matter sub 10 um) in excess of 10 times hazardous levels (Nguyen & Bullen, 2019). During these periods of intense smoke cover, Sydney had the worse air quality in the world (Clark, 2019)

Exposure to PM10 at elevated levels can cause numerous health hazards. It has been linked to (insert conditions). It can also exacerbate existing medical conditions, most commonly affecting those with asthma and/or other respiratory ailments (World Health Organisation, 2021). The cost to the healthcare system from exposure to PM10 is estimated to be as high as 8.4 billion dollars annually (Department of Environment and Conservation, 2005)

Air pollution is monitored through various meteorological sites across Sydney and NSW, most commonly by the NSW department of planning, industry and the environment (DPIE) which have over 20 sites within the Greater Sydney area (DPIE, 2021). Another project is the School Weather and Air Quality (SWAQ) which has an additional 6 air monitoring stations located in or around school campuses (SWAQ, 2021). Additional stations can be found online such as the citizen science project, World Air Quality Index Project, which has approximately 10 additional air monitoring sites across Greater Sydney, however they don't have an easy method of accessing raw data collection (WAQIP, 2021).

It would be beneficial in we were able to get an idea of the PM10 levels across Sydney at any give time. This would allow for tracking of smoke plums during bush fire season and generation of PM10 through industry sources. To do this we could substantially increase the number of air monitoring sites, building them at regularly spaced locations in and around Sydney, however this is unfeasible due to cost, expertise needed as well as maintenance. An alternative is to interpolate the values using the known values at known locations to predict unknown values across monitored regions of Sydney and potentially, larger areas of NSW. The most common forms of geospatial interpolation are Inverse Weight Distancing (IDW) which is computationally straightforward and Kriging which generally gives superior results but is significantly more computationally expensive to run.

Datasets

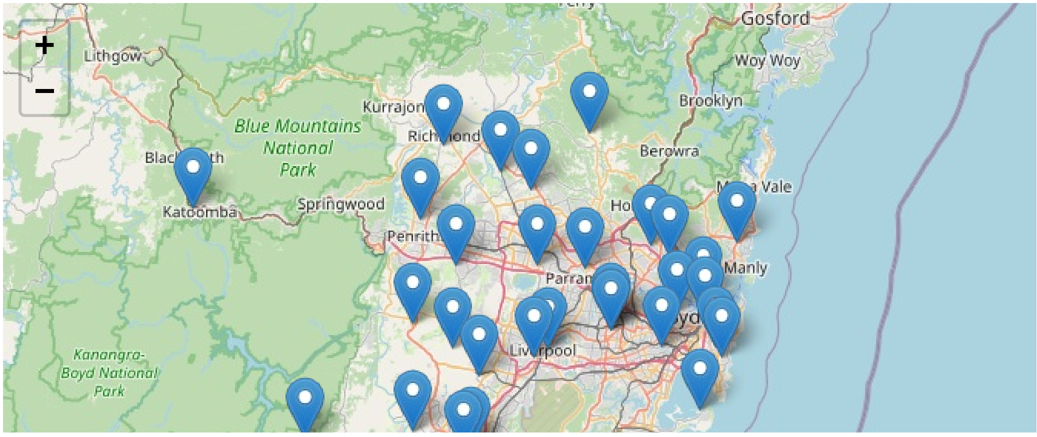
Two datasets have been used in this analysis. The first is the Air Quality Index (AQI) dataset. Values for a range of pollutants were captured every hour for 2019 and 2020, across all possible stations and downloaded using the AQI API interface. The stations used in the analysis were subsequently filtered to those within the Greater Sydney region. The latitude and longitude of the stations was stored in another file which was merged to contain all relevant information within the single file.

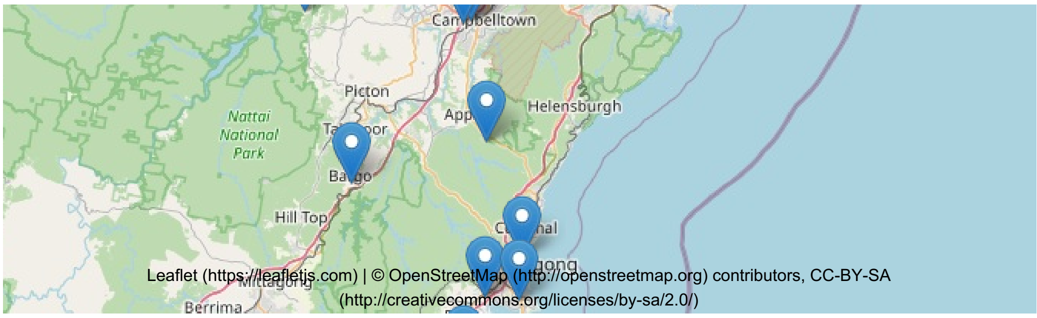
The second dataset is that of the SWAQ dataset. Various pollutant levels are recorded every 20 minutes, for our purpose, only values on the hour were collected. The latitude and longitude were only detailed on the SQAQ website and were joined manually for each station. Both the AQI and the SWAQ datasets were combined into a single purpose database.

Methods

The time frame for this analysis was chosen to be at 12pm on December 25th 2019. This was chosen as it's a date that a significant portion of people would be engaged in family activities in outdoor venues. It was also a day where Sydney wasn't completely engulfed in smoke allowing for larger variation in PM10 across the city. Any arbitrary date or hour could be chosen by modifying the 'Date' or 'Hour' option where date is in the format YYYY-MM-DD and hour can be between 1 and 24.

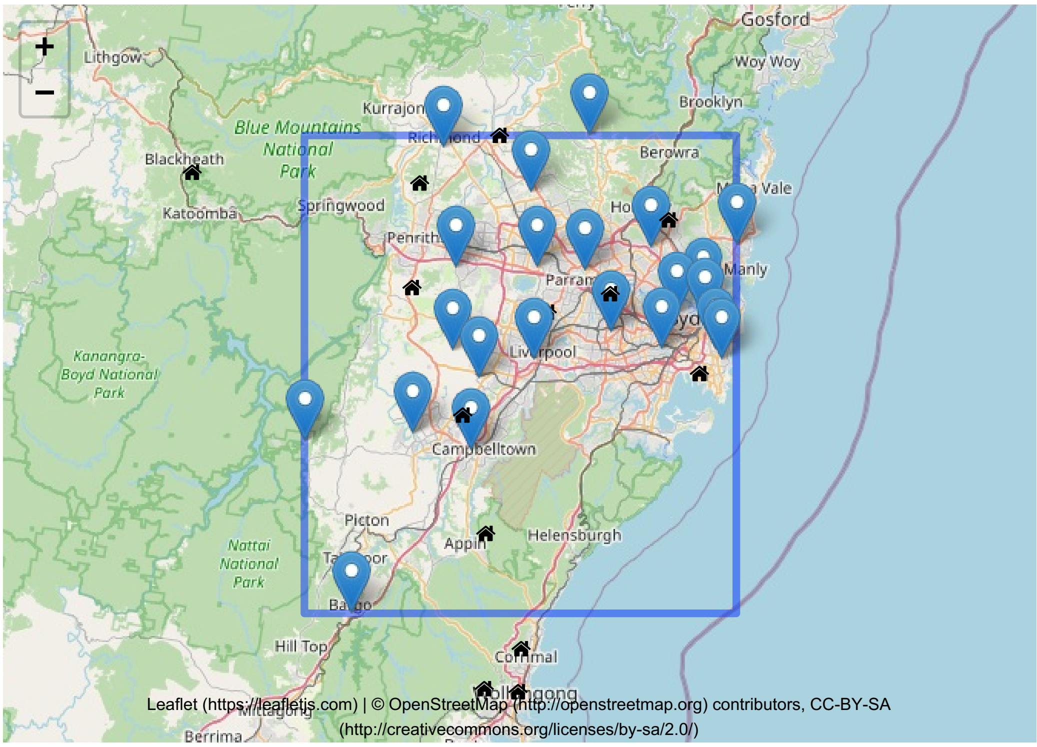
A dataset containing PM10 values from stations within the Greater Sydney region (fig 1) was generated. Further filtering to only numeric results and again filtering by values above 0 from these stations, only the stations shown in blue are used in the analysis while removed stations appear as red in fig 2. Stations record 'NA' when they were not operating. Due to instrument malfunction record values less than 0. These are not addressed earlier as due to the highly selective process of choosing a date and hour, it's unlikely an invalid numerical result will occur.





Leaflet (https://leafletjs.com) | © OpenStreetMap (http://openstreetmap.org) contributors, CC-BY-SA (http://creativecommons.org/licenses/by-sa/2.0/)

Figure 1. Initial selection of monitoring stations within Greater Sydney region



Leaflet (https://leafletjs.com) | © OpenStreetMap (http://openstreetmap.org) contributors, CC-BY-SA (http://creativecommons.org/licenses/by-sa/2.0/)

Figure 2. Map of filtered air monitoring stations in blue, stations with no or unrealistic data are in red

The empirical variogram is constructed from the dataset using the gstat package after the coordinates were modified to a spatial coordinates. The variogram displays the variation of PM10 recorded by various stations as a function of distance. The semi-variance measure is defined as half the average squared difference between various points separated by a distance. A variogram is able to display potential autocorrelation of the underlying stochastic process (MacKenzie et al, 2018).

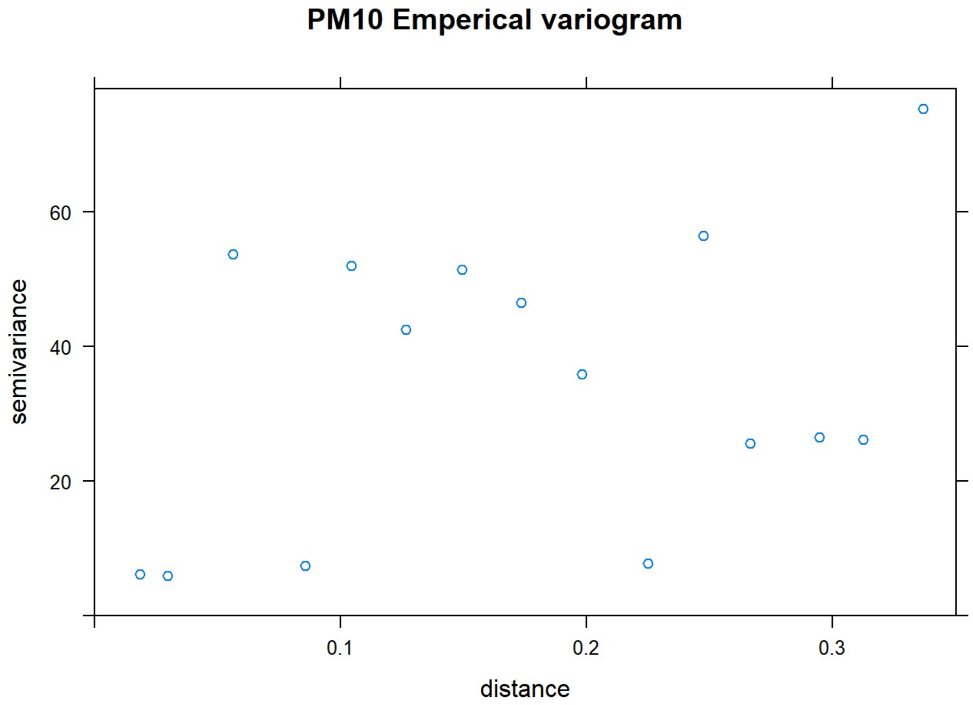


Figure 3. Emperical variogram of PM10

At this point we want to fit a model (nugget, model type, range, partial sill). Manually fitting a model can be intensive as the base packages in R such as gstat must converge in 200 iterations where small deviations from valid results can lead to the fit failing to converge. The automap package automates this process and tries to fit various model types. In this case, it modeled over 20 different models. The returned model is the one with the smallest residual sum of squares.

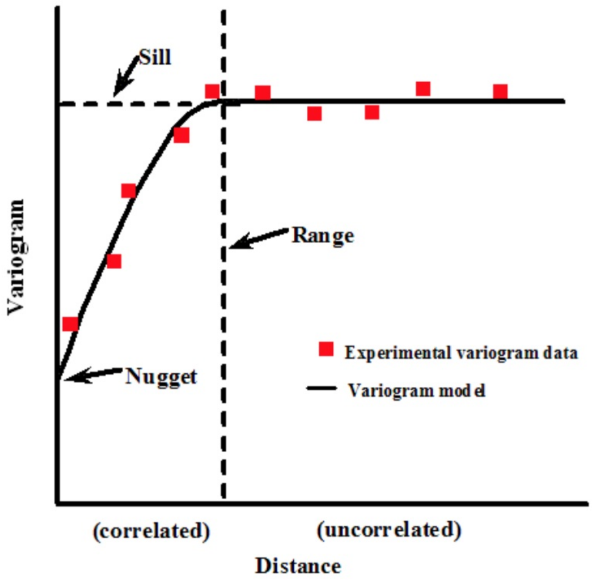


Figure 4. Example of an idealised variogram long with various important model parameter meanings

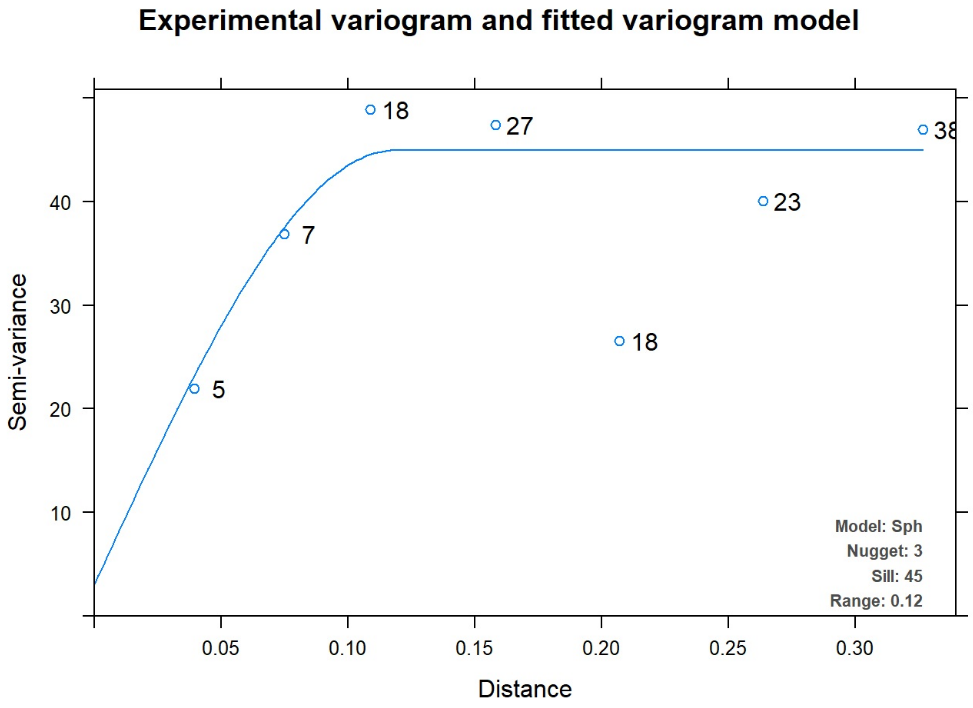


Figure 5. Plot generated by automap package using ordinary kriging using auotfitVariogram on data

We can see the result and see that it fits the empirical variogram reasonably well.

```
##   model     psill     range
## 1   Nug  3.042019 0.0000000
## 2   Sph 41.884181 0.1183857
```

Now that we have a variogram to the data with the model parameters (3.042, ‘Sph’,41.844,0.118), we can go ahead and create a grid. The vertical and horizontal lengths of the bounding box have been calculated using the haversine formula with the final result in meters. The final grid points are approximately 1km x 1km in size.

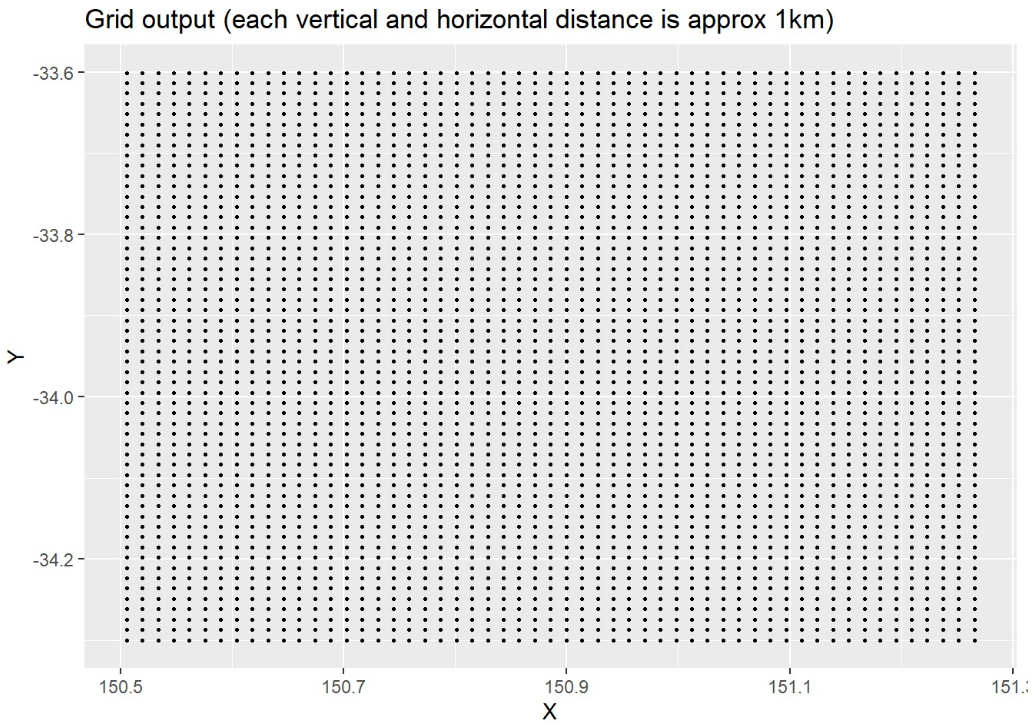


Figure 6. Generated grid that willl be used for Kriging

Results and Discussion

Before we do anything further, lets have a look at our initial data below in figure 7. The colour of each dot at each location is dependent on the PM10 concentration. Unfortunately, I was unable to work out a way to display the results over a map of Sydney.

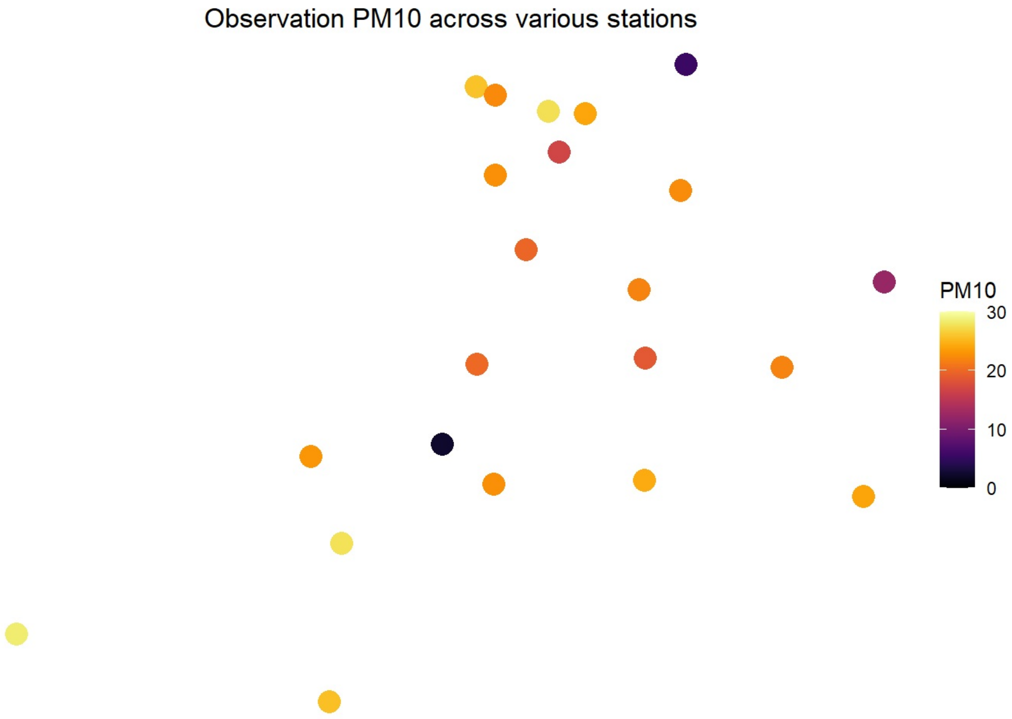


Figure 7. Initial look at all the air pollution stations

Ordinary Kriging

The first spatial interpolation method we will employ will be Ordinary Kriging where the variation in the grid field is a deviation from a constant average across the entire grid (Cressie, 1988). It results in the best linear unbiased prediction. The linearity occurs due to the estimation coming from a weighted combination of all available locations. The method tries to set the mean residual error to 0 meaning its unbiased and it also tries to minimize error variance.

Looking at the plot below (fig. 8), it appears there is an evening out of the PM10 values towards the borders of the grid, which is typically the farthest distance from a station location. The large yellow spots are attributed to the low PM10 concentrations observable in figure 7 as dark red circles.

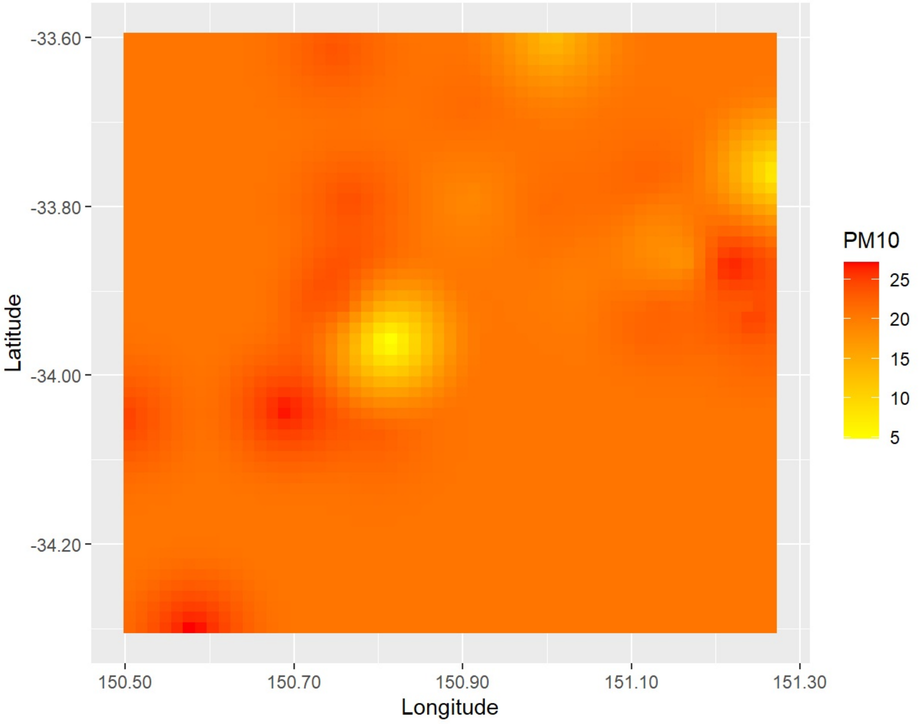


Figure 8. Ordinary Kriging output

The default ordinary kriging model assumes the data is anisotropic, that there is no directional effect. We can investigate the potential variation due to direction based on bearings where 000 is north, 090 is east and so on. The impact of directionality may be due to a number of factors such as geological and weather features or effects such as mountain ranges, coastline, wind speed and direction, population density etc. The below plot looks at north, east, south, and west directions at 0,90,180 and 270 respectively. We can see very little variation leading to the assumption that our data is anisotropic. This may be due to fire season where there are multiple sources of PM10 generation leading to potentially even coverage of PM10 at the levels recorded. It also might be that the number of locations is too small to pick up any variation. There is a slight difference in the N-S plots(0,180) compared to the E-W (090,270) direction.

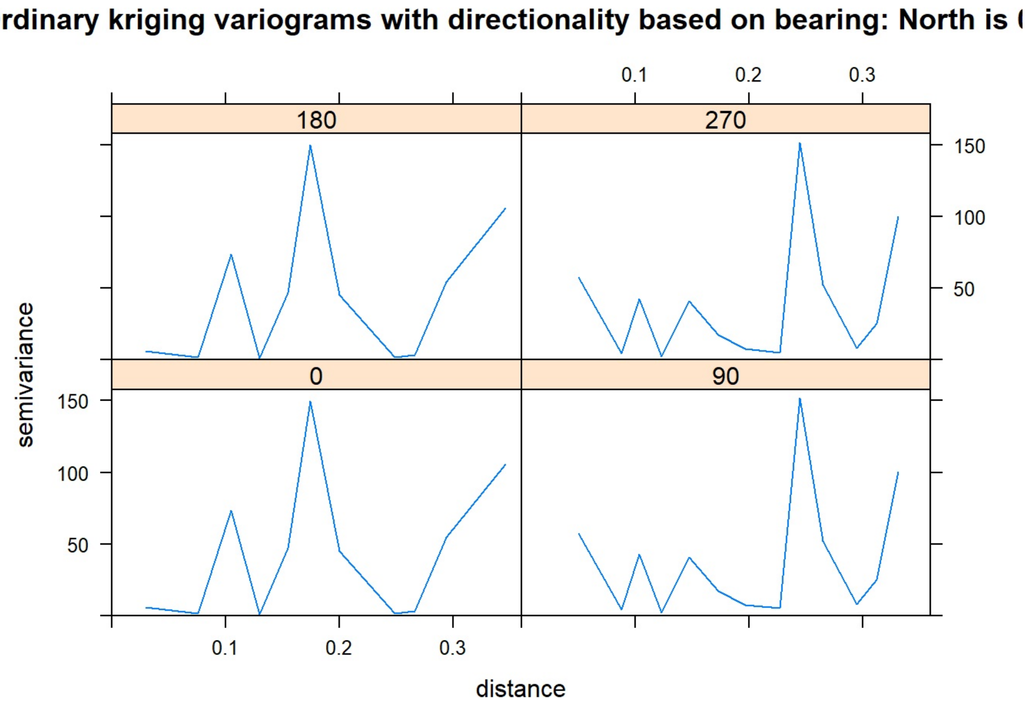


Figure 9. Ordinary Kriging: Assessing directionality

We can further explore the potential directionality by creating variograms for each direction and fitting a model. Again we use the bearing as input to create 4 plots. The model used will be the spherical model. From the figure 10. below, we can see that the distribution of the variogram data points are similar and that the spherical model appears similar for each direction. We will proceed with the assumption that the data is anisotropic.

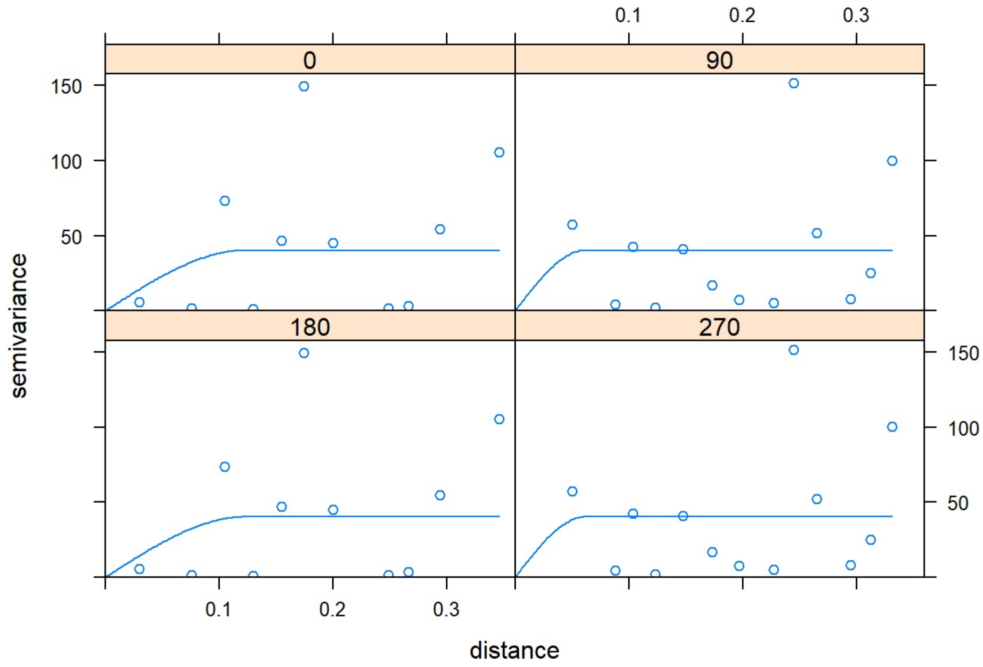


Figure 10. Ordinary Kriging Leave one out cross validation

Leave one out cross validation (LOOCV) is performed as there are too few data points to perform K fold cross validation. This removes one station and predicts values for its location which we can then compare to the real value and see how it performed.

The results of the LOOVC are shown below along with a summary output. The residuals can be viewed as a bubble plot where the colour is indicative of either positive or negative sign and the size is proportional to the size of the residual value at each location. We can see that the largest errors are correlated to the stations in figure 7 which were significantly lower compared to surrounding stations.

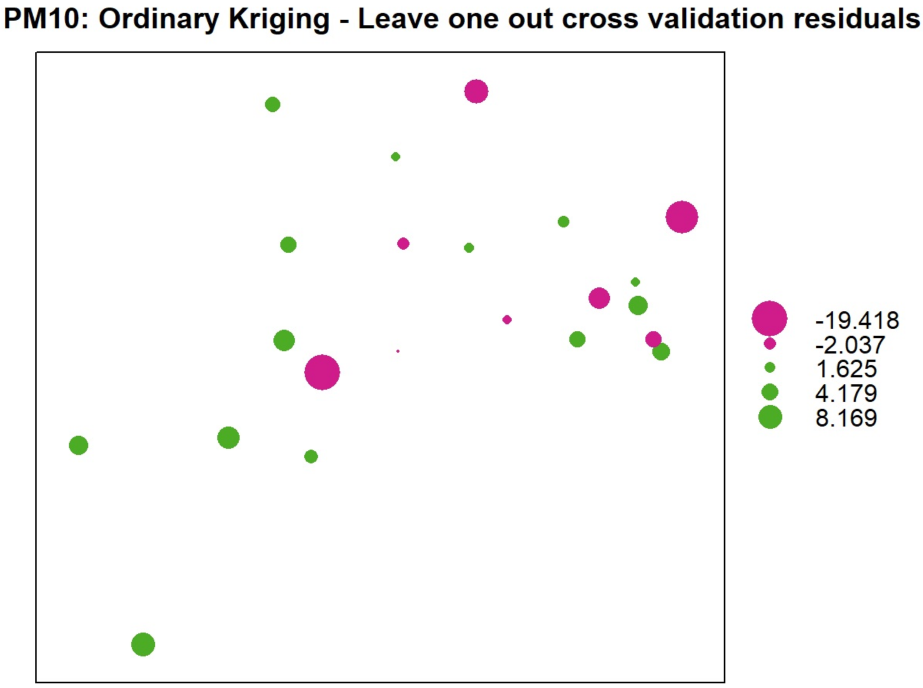


Figure 11. Ordinary Kriging Leave one out cross validation

Applying the autoKrige function to the data, we can generate a map of the predictions at various locations along with a plot of standard error and the fitted variogram. The automated process produces a similar plot seen in fig 9. where there is an trend towards homogeneous PM10 values at the edges.

```
## [using ordinary kriging]
```

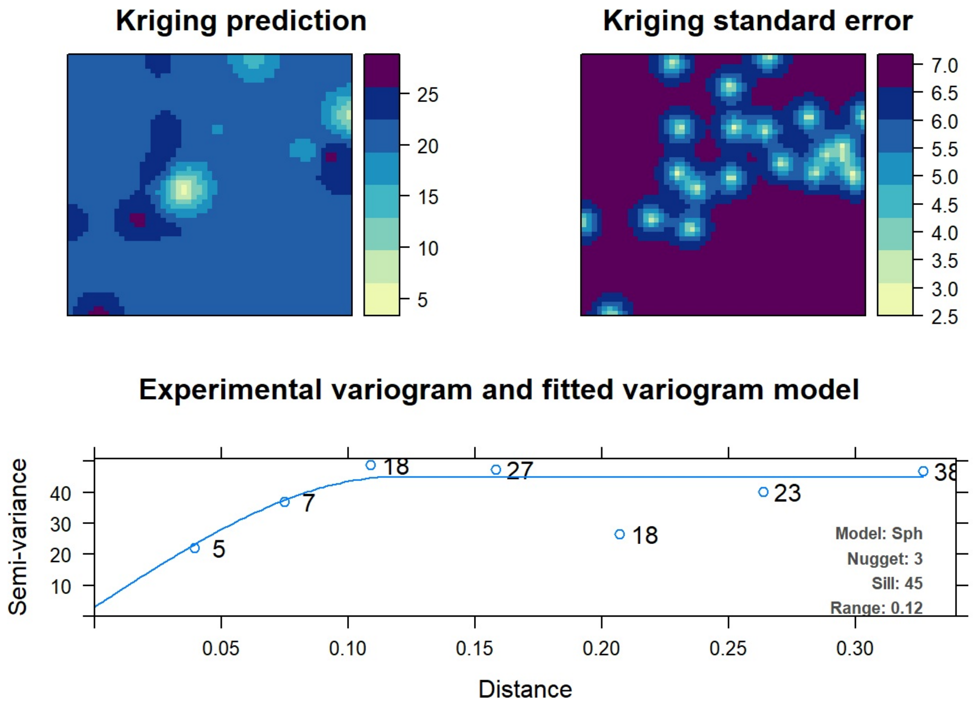


Figure 12. Ordinary Kriging autoKrige output

Due to the potential of outliers in the initial analysis which may have caused errors, we will also perform the analysis with outliers removed. The method used for outlier removal will be the interquartile range method.

After the outliers were removed we are down from 22 stations to 19 stations. Next we will create a new variogram and again let autofitVariogram fit the optimal model. We can see that the model fails to capture the data point fitting a linear model to a clearly parabolic shape. Unfortunately there are no default variograms that reasonably match the data output as they tend to follow an inverted shape to the one seen in the figure. As such we will proceed with the original dataset with the outliers included.

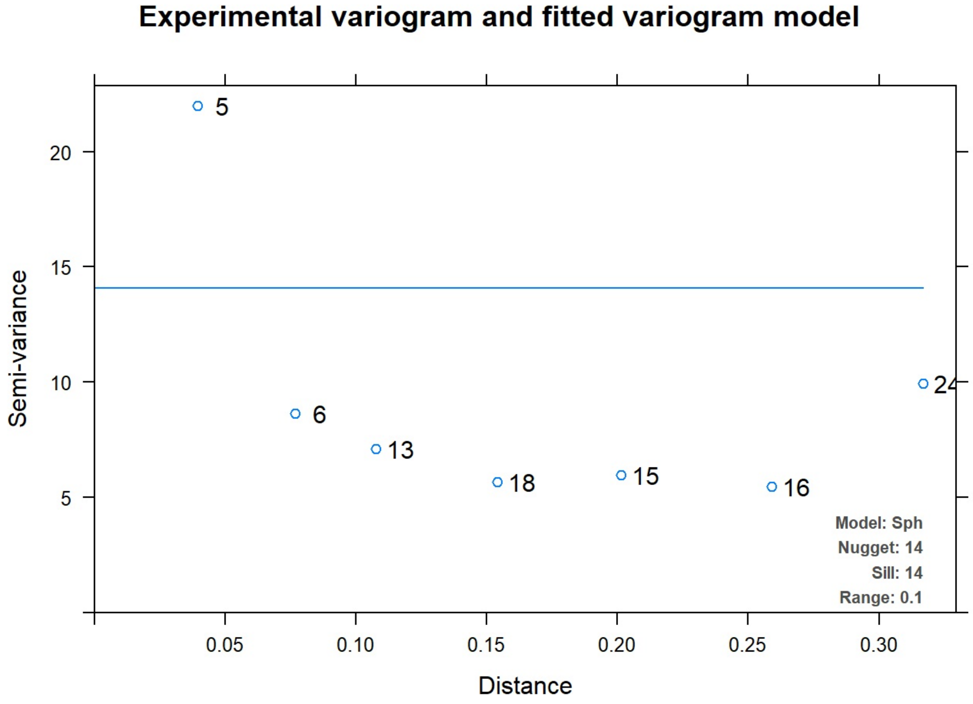


Figure 13. Outliers removed autofit variogram

Universal Kriging

Next we will try universal kriging using the autofitVariogram. This differs to ordinary kriging in that the mean is not considered equal but has a functional dependence on location. The terminology of drift is used which captures the tendency for the values to change as a function of the location (Kis, 2016).

Plotting the output of universal kriging, we get the following image (fig 14). It initially looks very similar to the plot from ordinary kriging (fig 8.) Looking closely there is a slight difference in the bottom right section where the results tend to follow a smooth gradient in comparison to the observed 3 red points.

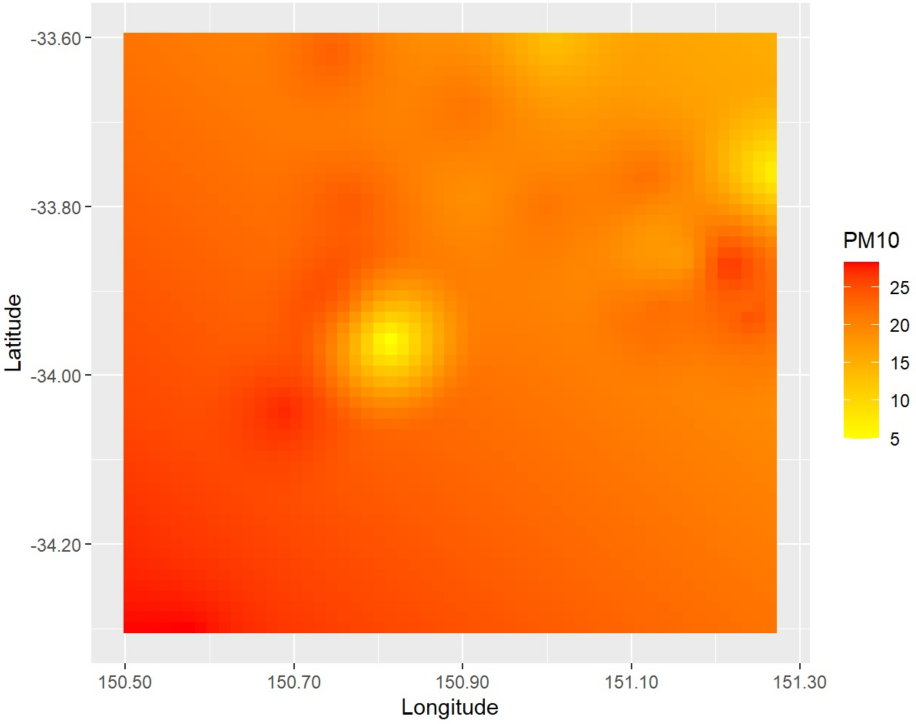


Figure 14. Universal kriging prediction map

Again applying the autoKrige function we can generate a prediction and standard error map. These image (figure 15.) do show a obvious distinction in comparison to the results in figure 12. The main difference just from eyeballing the figures, is that the main difference is that the errors are look less like point sources and the gradient tends to even out as we move away from the stations around the center.

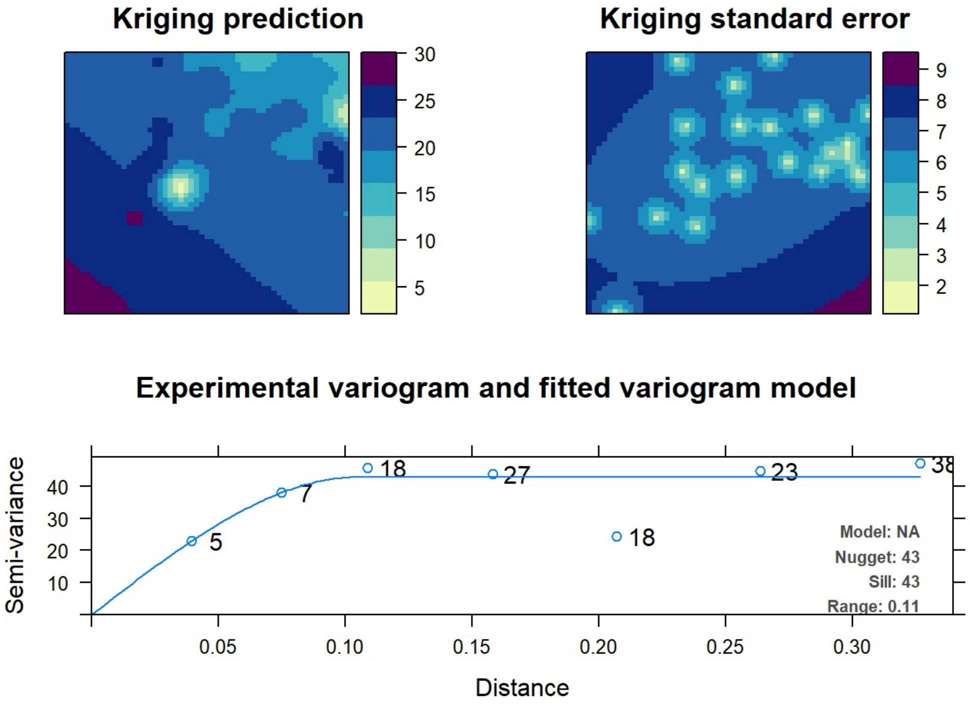


Figure 15. Universal kriging autoKrige output

```
#summary(universal_krige_leave_one_out_cv$observed)
universal_kriging_mean<-mean(universal_krige_leave_one_out_cv$residual) # mean error, ideally 0:
universal_kriging_rmse<-sqrt(mean(universal_krige_leave_one_out_cv$residual^2)) # RMSE, ideally small
universal_kriging_mnse<-mean(universal_krige_leave_one_out_cv$zscore^2) # Mean square normalized error, ideally c
lose to 1
```

Inverse Distance Weighting (IDW)

Finally Inverse Distance Weighting is used and relies upon deterministic assumptions. IDW is a much simpler method to implement compared to kriging algorithms. It differs in that it relies on the assumption that locations that are close in distance should be similar compared to those that are distant. Each predicted location is calculated using the nearest stations to make a prediction putting more emphasis on stations that are nearby and less emphasis on distant stations, hence the name.

```
## [inverse distance weighted interpolation]
```

```
idw_model %>% as.data.frame %>%
  ggplot(aes(x=coords.x1, y=coords.x2)) + geom_tile(aes(fill=var1.pred)) + coord_equal() +
  scale_fill_gradient(low = "yellow", high="red") +
  scale_x_continuous(labels=comma) + scale_y_continuous(labels=comma) +
      xlab("Longitude") + ylab("Latitude")+labs(fill = "PM10")
```

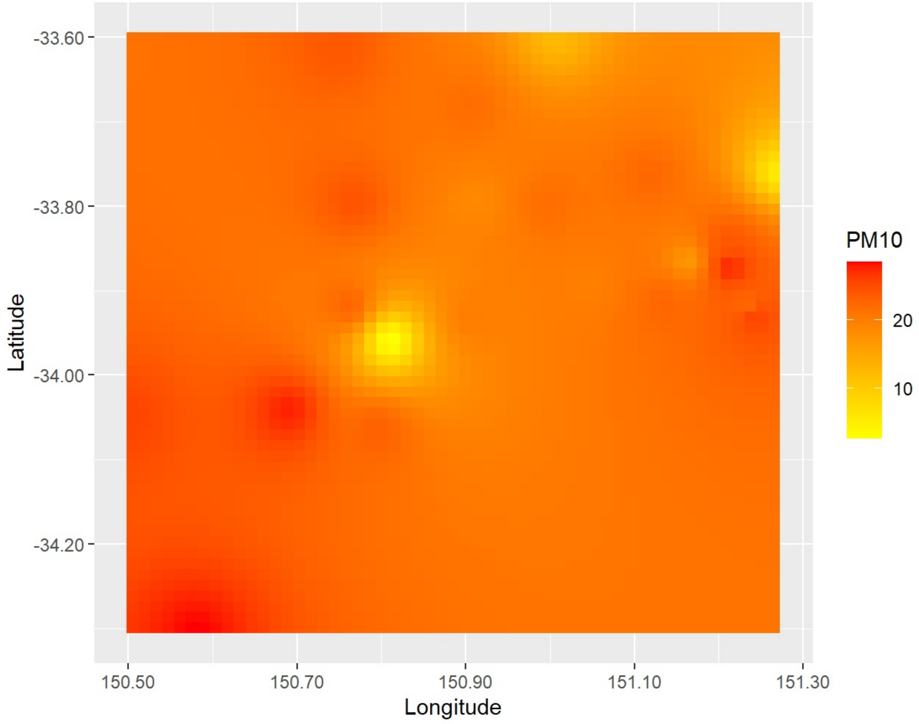


Figure 17. Inverse Distance Weighting heatmap

We an also perform LOOCV on the inverse distance weighting model to find the RMSE of 6.87

We will compare the RMSE of both types of kriging and the IDW method to determine which method was the most accurate. The RMSE of ordinary kriging and universal kriging was 7.00 while for IDW, it was 6.87. It is quite strange that both kriging results give the same error while producing different standard prediction and standard error plots (fig 11 and 14). The errors do change when the grid size is altered with Universal kriging recording a slightly lower RMSE of 6.89 when grid size is approx 100m x 100m although the computation takes a considerable while longer. Typically kriging methods will outperform IDW methods (Gong et al, 2014) however there is also evidence of IDW outpeforming kriging and other interpolation methods (Zarco-Perello & Simoes, 2017). If we had chosen a different hour and date or if we had access to more data, we might see that result. It might also be that this is an example of the no free lunch theorem.

A problem with geospatial interpolation is that the station location may be assumed to be the source of PM10 which is not the case. This can explain what appear as point sources in figures 12 and 15. It may also be the case that the recorded value of the stations is the highest potential value which is likely wrong again if the station is far from the source of pollution. One such scenario may be that there is a PM10 source and lets say 1km east and west of the location are air monitoring stations. Various algorithms may assume the stations are point sources and assume that the PM10 concentration may drop between the sources when in this case, that is the opposite of what's happening.

We can see from this work and other similar bodies of work that a high number of location measurements are necessary to build a usable interpolation model. The number of stations available in this work is likely inadequate to draw significant conclusions on.

Future work

While kriging is generally accepted to be one of the best interpolation methods for geospatial data, other simpler methods such as thin plate spline regression, nearest neighbours which is very similar to another method using thessian polygons. In addition there are also more advanced methods available such as land-use regression. Other kriging models such as cokriging may also improve the model where we can use multiple variables to krige, in our case, wind speed and direction may improve the model. Final complicatedly we could also add a temporal element, kriging over time. This has been attempted in one of the github Rmd files (github.com/DStockdale1/Sydney_PM10_Kriging/ spatial_temporal analysis.Rmd) but was too complicated along with a lack of documentation made progress very slow.

Conclusion

Three models of spatial interpolation have been used to create a plot of particulate matter sub 10um (PM10) over the Greater Sydney region using the multiple air quality stations within the area. Due to the small number of air quality stations, the number of data points is low and may be reduced further due to instrument maintenance or malfunction at any given time. The first model uses ordinary kriging, the second uses universal kriging and the final model implements universal distance weighting

Acknowledgements

https://swilke-geoscience.net/post/2020-09-10-kriging_with_r/kriging/ (https://swilke-geoscience.net/post/2020-09-10-kriging_with_r/kriging/) https://rpubs.com/nabilabd/118172 (https://rpubs.com/nabilabd/118172) https://rpubs.com/hungle510/202760 (https://rpubs.com/hungle510/202760) https://rstudio-pubs-static.s3.amazonaws.com/46259_d328295794034414944deea60552a942.html

(https://rstudio-pubs-static.s3.amazonaws.com/46259_d3282957940344149 44deea60552a942.html)

Appendix

```
## Checking if any bins have less than 5 points, merging bins when necessary...
## 
## [[1]]
##   model     psill      range kappa
## 1   Nug  9.191644  0.0000000     0
## 2   Ste 28.214949 -0.0277017     5
## 
## [[2]]
##   model    psill       range kappa
## 1   Nug 10.77520  0.00000000     0
## 2   Ste 26.49722 -0.02124451    10
## 
## ^^^ ABOVE MODELS WERE REMOVED ^^^
## 
## Selected:
##   model     psill     range
## 1   Nug  3.042019 0.0000000
## 2   Sph 41.884181 0.1183857
## 
## Tested models, best first:
##    Tested.models kappa    SSerror
## 1            Sph     0   191243.4
## 23           Ste     2   211870.2
## 22           Ste   1.9   213126.6
## 21           Ste   1.8   214565.2
## 20           Ste   1.7   216221.4
## 19           Ste   1.6   218139.4
## 18           Ste   1.5   220375.2
## 17           Ste   1.4   223000.9
## 16           Ste   1.3   226110.4
## 15           Ste   1.2   229827.3
## 14           Ste   1.1   234318.9
## 13           Ste     1   239815.1
## 12           Ste   0.9   246640.4
## 11           Ste   0.8   255266.1
## 10           Ste   0.7   266402.1
## 9            Ste   0.6   281165.1
## 8            Ste   0.5   301413.3
## 2            Exp     0   301413.5
## 7            Ste   0.4   330472.5
## 6            Ste   0.3   374929.2
## 5            Ste   0.2   450002.1
## 3            Gau     0   762188.0
## 4            Ste  0.05 43259127.4
```

References

Clark, G., (2019), Sydney’s air quality worst in the world due to bushfires, Daily Telegraph, accessed on Novemebr 11 2021 from https://www.dailytelegraph.com.au/news/nsw/sydneys-air-quality-among-worst-in-the-world-due-to-bushfires/news-story/0c016c0575860fc371605542435832ad (https://www.dailytelegraph.com.au/news/nsw/sydneys-air-quality-among-worst-in-the-world-due-to-bushfires/news-story/0c016c0575860fc371605542435832ad)

Cressie, N., (1998), Spatial prediction and ordinary kriging, Mathematical Geology, 20, 405–421,DOI https://doi.org/10.1007/BF00892986 (https://doi.org/10.1007/BF00892986)

Department of Environment and Conservation, (2005), Air Pollution Economics - Health Costs of Air Pollution in the Greater Sydney Metropolitan Region

Department of Planning Industry and the Environment, (2021), Air quality map - Sydney, News South Wales Department of Planning, Industry and Environment, accessed on November 14 from https://www.dpie.nsw.gov.au/air-quality/air-quality-maps/sydney-map (https://www.dpie.nsw.gov.au/air-quality/air-quality-maps/sydney-map)

Fire and Rescue NSW,Fire and Rescue New South Wales Annual Report 2019-2020, (2020), accessed on Novemeber 14 2021 from https://www.fire.nsw.gov.au/gallery/files/pdf/annual_reports/annual_report_2019_20.pdf (https://www.fire.nsw.gov.au/gallery/files/pdf/annual_reports/annual_report_2019_20.pdf)

Gong, G., Mattevada, S., O’Bryant, S.E., Comparison of the accuracy of kriging and IDW interpolations in estimating groundwater arsenic concentrations in Texas, Environmental Research, vol 30 pg 59-

Kis, I,M., Comparison of Ordinary and Universal Kriging interpolation techniques on a depth variable (a case of linear spatial trend), case study of the Šandrovac Field, The Mining-Geology-Petroleum Engineering Bulletin, 10.17794/rgn.2016.2.4

Li, J., Heap, A.D., (2017), A Review of Spatial Interpolation Methods for Environmental Scientists, accessed November 14 2021 from https://data.gov.au/data/dataset/a-review-of-spatial-interpolation-methods-for-environmental-scientists (https://data.gov.au/data/dataset/a-review-of-spatial-interpolation-methods-for-environmental-scientists)

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L., Hines, J.E., Chapter 4 - Basic Presence/Absence Situation, Occupancy Estimation and Modeling (Second edition), Inferring Patterns and Dynamics of Species Occurrence, DOI 10.1016/B978-0-12-407197-1.00006-5

Nguyen, K., Bullen, J., (2019), Sydney smoke three times worse this NSW bushfire season, but health effects from ‘medium-term’ exposure unclear, ABC news, accessed on Novemeber 14 2021 from https://www.abc.net.au/news/2019-12-03/sydney-air-quality-smoke-haze-worse-this-bushfire-season/11755546 (https://www.abc.net.au/news/2019-12-03/sydney-air-quality-smoke-haze-worse-this-bushfire-season/11755546)

SWAQ, (2021), Empowering urban weather research with schools, accessed November 14 from https://www.swaq.org.au/ (https://www.swaq.org.au/)

WAQIP, 2021, Air Pollution in Sydney: Real-time Air Quality Index Visual Map, accessed on November 14 2021 from https://aqicn.org/map/sydney/ (https://aqicn.org/map/sydney/)

World Helath Organisation (WHO), (2021), Ambient (outdoor) air pollution, accessed on NOvemeber 14 2021 from who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Zarco-Perello, S., and Simões, N. (2014). Ordinary Kriging vs inverse distance weighting: spatial interpolation of the sessile community of Madagascar reef, Gulf of Mexico.