**Statistical Analysis of Calgary’s Historical Air Quality Data**

The domain that we will be working with for this project is Environment and safety. Within this domain, we are trying to investigate the patterns in air quality data in Calgary over 39 years. The analysis of this data will allow us to draw conclusions about the overall air quality status in Calgary, how much on average of each major greenhouse gasses are being produced on a daily basis and how this data compares to the global data yearly. This will show us if Calgary is doing better or worse than the rest of the world and give us an idea of the precautions humans should take to protect our respiratory health. It will also help us take responsible eco-friendly actions to protect our environment especially due to climate change. The population for this data would be 158,532 which is the number of rows in our dataset and the variables we will be using are “Date” and “Parameter”. Within each parameter, we are interested in the daily average values of Carbon Dioxide, Methane, Nitrogen Dioxide, Nitric Oxide and PM2.5 mass which were recorded using various methods such as “Instrumental”, “Calculated”, “Sharp” etc. at 5 different locations in Calgary. For the global data, the variables we will use are PM2.5 and Year.

The datasets we have chosen for our project is “Historical Air Quality” which was collected by the Calgary Region Airshed Zone and submitted to Alberta Environment and Parks (AEP) and WHO Air Quality Database. This information is publicly available and can be used from the City of Calgary’s Open Data Portal and WHO website. These datasets are in a structural tabular format. The visualizations we will use to visually inspect the data are density plots to visualize how greenhouse gasses changed over 39 years as well as line/scatter plots to visualize the PM2.5 mass of global data vs Calgary data over the years that overlap between the two datasets.

The focus of our statistical investigation is to understand if there is any statistical difference between the particulate matter recorded in Calgary vs particulate matter recorded globally in order to gain evidence to support or reject our hypothesis that PM2.5 globally is equal to PM2.5 in Calgary. We also want to determine how the greenhouse gas emissions changed over the years in Calgary by plotting the distribution of greenhouse gas emissions over the years and estimating the population average via bootstrapping. This data is leveraged by the Government of Alberta to visualize, analyze and continually monitor air pollution. This enables Calgary residents to protect themselves and take the necessary precautions especially those who are health compromised.

The first statistical method we will use is bootstrap statistics to estimate population mean of the major greenhouse gasses such as Carbon Dioxide, Methane, Nitrogen Dioxide and Nitric Oxide. Since the recommended sample size is 10% of population size as long as it does not exceed 1000, we will be using 1000 as our sample size to randomly take samples and simulate 3000 times and calculate lower bound and upper bound via bootstrap confidence interval to estimate the population mean of the greenhouse gasses yearly. The second statistical method we will be using is Hypothesis testing- Single Parameter (two populations). We will be using this method to check if there is any difference between the PM2.5 Mass (Particulate matter of width 2.5 microns) observed in Calgary vs. globally. Our null hypothesis would be that there is no difference between Calgary and global data and our alternative would be that global PM2.5 is greater than that of Calgary’s. We will also use bootstrap or conventional technique to create confidence intervals based on the normality of the data being used.

**References**

City of Calgary’s Open Data Portal *Historical Air Quality* [online]. Available at:<https://data.calgary.ca/Environment/Historical-Air-Quality/uqjm-jxgp/data>

WHO Air Quality Database (Update 2022), Available at:https://www.who.int/data/gho/data/themes/air-pollution/who-air-quality-database

Denchak, Melissa. “Greenhouse Effect 101.” NRDC, 16 July 2019, www.nrdc.org/stories/greenhouse-effect-101#:~:text=Bednarski%20for%20NRDC-.

Steven R.Hanna, 1989. Confidence limits for air quality model evaluations, as estimated by bootstrap and jackknife resampling methods. Available at: <https://www.sciencedirect.com/science/article/pii/0004698189901613> [Accessed October 4, 2022].

Lee-Ing Tong, Chih-Wei Chang, Shin-En Jin, R.Saminathan, 2012. Quantifying uncertainty of emission estimates in National Greenhouse Gas Inventories using bootstrap confidence intervals. Available at: <https://www.sciencedirect.com/science/article/pii/S1352231012003044> [Accessed October 4, 2022].

W. GeoffreyCobourn, 2010. An enhanced PM2.5 air quality forecast model based on nonlinear regression and back-trajectory concentrations. Available at: <https://www.sciencedirect.com/science/article/pii/S1352231010003821> [Accessed October 4, 2022].

Christopher Kennedy, Julia Steinberger, Barrie Gasson, Yvonne Hansen, Timothy Hillman, Miroslav Havránek, Diane Pataki, Aumnad Phdungsilp, Anu Ramaswami, and Gara Villalba Mendez, 2009. Greenhouse Gas Emissions from Global Cities. Available at: <https://pubs.acs.org/doi/full/10.1021/es900213p> [Accessed October 4, 2022].

Daniel Hoornweg, Lorraine Sugar, and Claudia Lorena Trejos Gómez, 2011. Cities and greenhouse gas emissions: moving forward. Available at: <https://journals.sagepub.com/doi/abs/10.1177/0956247810392270> [Accessed October 4, 2022].

J. Jason West, Steven J. Smith, Raquel A. Silva, Vaishali Naik, Yuqiang Zhang, Zachariah Adelman, Meridith M. Fry, Susan Anenberg, Larry W. Horowitz & Jean-Francois Lamarque, 2013. Co-benefits of mitigating global greenhouse gas emissions for future air quality and human health. Available at: <https://www.nature.com/articles/nclimate2009> [Accessed October 4, 2022].

Tome Eftimov & Peter Korošec, Chapter 4, Introduction to Statistical Analysis. Available at: <https://link.springer.com/chapter/10.1007/978-3-030-96917-2_4> and [**https://rdcu.be/cWTZx**](https://rdcu.be/cWTZx)[Accessed October 4, 2022].

Paul S Levy and Kristine Stolte, 2000. Statistical methods in public health and epidemiology: a look at the recent past and projections for the next decade. Available at: <https://journals.sagepub.com/doi/abs/10.1177/096228020000900106> [Accessed October 4, 2022].