WHICH REGIONS IN SINGAPORE HAVE RELATIVELY POORER AIR QUALITY

Introduction

Pollutant Standard Index (PSI) in Singapore is typically used as an indicator of air quality, and is calculated based on the levels of PM2.5, PM10, ozone, sulfur dioxide, nitrogen dioxide and carbon monoxide. Recently, climate change has been a persisting issue, with its causes contributing to poorer quality of air. Frequent burning of crops in Indonesia to clear land for agriculture, vehicle emissions, wildfires and more release large amounts of smoke which is then carried to Singapore via winds.

Purpose and motivation

Air quality has significant public health relevance for Singapore, particularly as the country transits into a super-aged society, where one in four citizens are aged 65 and above. Numerous studies have shown that older adults are physiologically more vulnerable to negative effects of air pollution due to age-related decline in respiratory and cardiovascular function, leading to increased prevalence of chronic diseases such as chronic obstructive pulmonary disease, asthma, and hypertension. As our parents, older relatives, grandparents and teachers continue to age and their health begins to decline, our group was motivated to find the best are for them to retire and live in. Since air quality is mostly consistent throughout the area, our group has decided to focus on air quality as our first deciding factor in finding the best area for retirement, by ruling out certain regions with the poorest air quality from our choices.

Research question

1. Do PSI levels vary significantly between regions?
2. Which regions have relatively poorer air quality?

Hypothesis

1. PSI does vary significantly between regions.
2. West has poorest air quality with highest daily PSI levels and highest extreme PSI levels.

Methodology

The hourly PSI level in each region (North, South, East, West and Central) for 2023 and 2024 were obtained from a dataset named “Historical 24h PSI” from data.gov.sg, which contains data recorded by the National Environmental Agency. PSI data from January 2023 to December 2024 was used for our calculations to ensure the data is relevant (within the last 3 years) and reliable. For ease of calculation, the monthly average PSI level for each of the 5 regions was calculated to be used for data analysis

Data Analysis

As more than 2 regions are being compared, Analysis of Variance (ANOVA) or Kruskal-Wallis test should be used instead of a t-test to determine if difference in PSI levels between regions is statistically significant. This minimizes type 1 error (when the null hypothesis is incorrectly rejected) which would increase when multiple t-tests are done. While ANOVA and Kruskal-Wallis both serve similar purposes, their methodologies and assumptions are different. ANOVA is a parametric test which assumes normality while Kruskal-Wallis is a non-parametric alternative to ANOVA that does not assume normality. Hence, in order to determine which test should be used, the Shapiro-Wilk test for normality should first be done.

1. Shapiro-Wilk test (for normality)

H_0: Dataset has a normal distribution
H_a: Dataset does not have a normal distribution

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where n is the number of average monthly PSI readings per region, a_i is the corresponding coefficient vector of the weights of the Shapiro-Wilk test (obtained from the Shapiro-Wilk test table), $x_{(i)}$: is the i^{th} order statistic and $\bar{x}$ is the mean of the dataset.

The W statistic represents the dataset’s resemblance to a normal distribution. The higher the W statistic, the closer the dataset follows a normal distribution.

Region	North	South	East	West	Central
W	0.9491	0.9567	0.9199	0.9275	0.9801

Figure 1: Table of calculated W statistic for each region

n\W	0.01	0.02	0.05	0.1	0.5	0.9	0.95	0.98	0.99
24	0.884	0.898	0.916	0.93	0.963	0.981	0.984	0.987	0.989

Figure 2: An exercept of the Shapiro-Wilk test table of p-values for n = 24

Comparing Tables 1 and 2, for n = 24, all W statistic values are above $W_{0.05}$ = 0.916, hence all p-values are above 0.05. Since p-value is greater than α = 0.05, the null hypothesis is not rejected and we can conclude that all the datasets follow a normal distribution.

2. Analysis of variance (ANOVA)

Since we have concluded that the datasets follow a normal distribution, we can use ANOVA (Single factor) to determine whether the difference in PSI levels between each region is statistically significant. It does this by evaluating the variability within groups and between groups.

H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: at least one of the means for a region is different.

Region	Count	Sum	Average	Variance
North	24	918.077	38.253	80.812
South	24	1035.478	43.145	40.064
East	24	1166.015	48.584	45.563
West	24	980.515	40.855	92.010
Central	24	1124.055	46.8356	83.189

Figure 3: Summary table for ANOVA (Single factor) of count, sum, average and variance per region

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1718.8522	4	429.7130	6.2890	0.0001	2.4506
Within Groups	7857.6872	115	68.3277			
Total	9576.5393	119				

Figure 4: Summary table for ANOVA (Single factor) of , sum, average and variance per region

Since F > F crit (6.2890>2.4506), we reject the null hypothesis so not all of the five means from each region are equal.

3. Tukey’s test

In order to compare the PSI between regions, we can perform pairwise comparisons between each region.

$$q_{tukey} = \frac{Absolute\ difference}{\sqrt{\frac{MS}{n}}}$$

where n is the number of average monthly PSI readings per region and MS is the (mean square) average of squared deviations.

Region	Absolute difference	Standard error ($\sqrt{(MS/n)}$)	q tukey	q critical value	Significant?
N vs S	4.892	1.687	2.899	3.917	No
N vs E	10.331	1.687	6.123	3.917	Yes
N vs W	2.602	1.687	1.542	3.917	No
N vs C	8.582	1.687	5.086	3.917	Yes
S vs E	5.439	1.687	3.224	3.917	No
S vs W	2.290	1.687	1.357	3.917	No
S vs C	3.691	1.687	2.187	3.917	No
E vs W	7.729	1.687	4.581	3.917	Yes
E vs C	1.748	1.687	1.036	3.917	No
W vs C	5.981	1.687	3.545	3.917	No

Figure 5: Table showing calculation and result of Tukey test

From table 4, degrees of freedom for denominator is 115, hence referring to the critical values of studentised range distribution (q) for familywise alpha =0.05, and given that number of regions is 5, critical value of q is 3.917
From table 4, degrees of freedom for denominator is 115, hence referring to the critical values of studentised range distribution (q) for familywise alpha =0.05, and given that number of regions is 5, critical value of q is 3.917.
When calculated q is greater than critical value of q, difference in mean PSI levels between the two regions is statistically significant.
Hence, East had a higher mean PSI than both North and West, while Central had a higher mean PSI than North

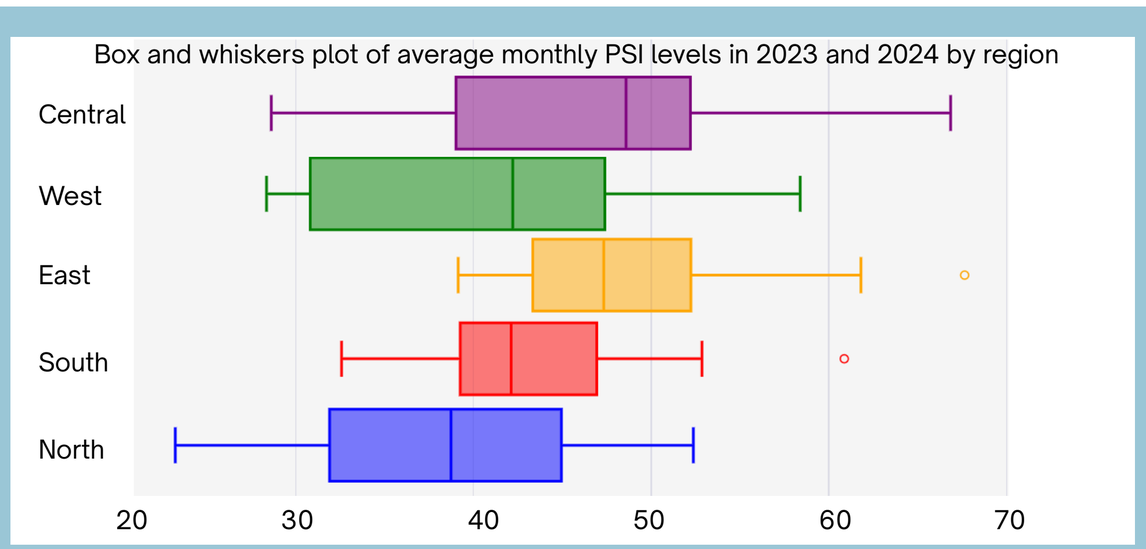

Figure 6: Graph of Box and whiskers plot of average monthly PSI levels in 2023 and 2024 by region

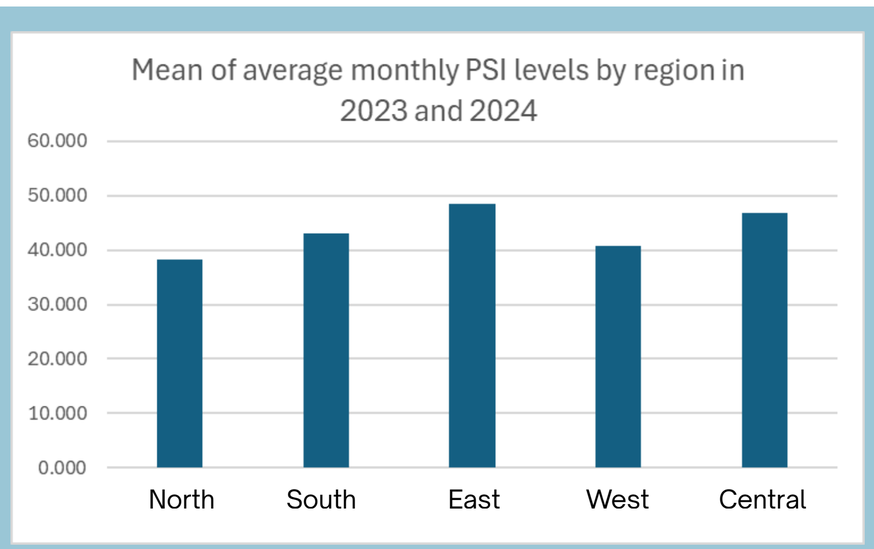

Figure 7: Bar Graph showing mean of average monthly PSI levels (by region) in 2023 and 2024.

From graph 1, the median PSI levels for north, south, east, west and central are 38.7, 42.1, 47.3, 42.2 and 48.6 respectively. Hence, central region has the highest median PSI level and largest PSI level range. So central has the greatest magnitude of fluctuation of PSI levels. Result of comparing median in box and whiskers plot and result of Tukey’s test comparing East and Central differ due to the mean for East being affected by an outlier of 67.6, which is the highest monthly average PSI level recorded for any region in 2023 to 2024. Hence, despite central region tending to have a higher daily PSI level due to its higher median, East likely experiences higher extreme PSI levels which could have more adverse and immediate effects on the health of seniors.

Conclusion

From Data analysis, since null hypothesis of ANOVA was rejected, this proves our first hypothesis as the PSI levels do vary significantly between regions. Next, from comparing of graphs and result of Tukey’s test, our second hypothesis is disproven as Central is likely to have the highest monthly PSI level while East is likely to have highest extreme PSI levels.
Based on the tests and graphs above, it can be seen that the East region in Singapore has the highest PSI levels, and hence poorest air quality. This may be due to a variety of reasons.
Firstly, wind patterns steer towards the East, both seasonally and daily. During the Northeast Monsoon (Dec–Mar), low-level winds blow from the NE, pushing regional pollution toward the SE/E corridors, blowing haze and smoke towards the East. As East commonly has stable evenings with high atmospheric stability and light winds, fine particles can be trapped near the surface, so the Eastern region which is downwind of local sources that evening can spike in PSI levels.
Secondly, aviation at Changi Airport concentrates aircraft NO_x/PM during take-off and airside activity, contributing to gas emissions in the region.
Additionally, the East has shipping lanes just offshore: The Singapore Strait’s heavy traffic emits NO_x, SO_2, and PM. On NE or E sea-breeze days, portions of wind can push toward the East and Southeast.
Lastly, coastal meteorology & recirculation involves sea-breeze fronts (frequent in inter-monsoon months) moving inland from the south/east coasts in the afternoon, potentially recirculating urban/port/airport emissions and raising $PM_{2.5}$/NO_2 near the East Coast.
The East’s dense residential communities along the coastal corridor plus proximity to Changi/Strait means higher PSI levels can be more consequential for older adults (cardiopulmonary vulnerability) compared to other regions in Singapore.

1.Science Talk: No fires, so why the haze? Why calm winds are bad for urban cities https://www.straitstimes.com/singapore/psi-hits-unhealthy-range-in-singapore-as-air-quality-worsens
2. Rynearson, S. (2018). Learn VBA Online - Tutorial for Beginners (Free and Interactive). Automateexcel.com; Automate Excel. https://www.automateexcel.com/learn-vba-tutorial/
3.Zaiontz, C. (n.d.). Shapiro-Wilk Table | Real Statistics Using Excel. Real-Statistics.com. https://real-statistics.com/statistics-tables/shapiro-wilk-table/
4.Shapiro-Wilk Test Excel and Google Sheets. (n.d.). Automate Excel. https://www.automateexcel.com/stats/shapiro-wilk-test/
5.Excel Easy. (2010). Anova in Excel. Excel-Easy.com. https://www.excel-easy.com/examples/anova.html
6.Macsin, A. (2022). Tukey Test / Tukey Procedure / Honest Significant Difference. Statistics How To. https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/post-hoc/tukey-test-honest-significant-difference/
7.Stephanie. (2017, July 26). Studentized Range Distribution. Statistics How To. https://www.statisticshowto.com/studentized-range-distribution/#qtable
8.Science Talk: No fires, so why the haze? Why calm winds are bad for urban cities https://www.straitstimes.com/singapore/environment/science-talk-no-fires-so-why-the-haze-why-calm-winds-are-bad-for-urban-cities