Air Pollution in Seoul

Achilleas Prokopakis, Elias Michael

Abstract— Air pollution poses a significant challenge in urban environments, affecting public health and environmental quality. This study investigates the dynamics of air pollution in Seoul, using a dataset containing hourly measurements of six key pollutants—SO_2, NO_2, CO, O_3, PM10, and PM2.5—collected from 25 monitoring stations between 2017 and 2019. The analysis focuses on uncovering spatiotemporal patterns, identifying pollution hotspots, and exploring pollutant diffusion. Temporal aggregation methods and visual tools such as line plots, heatmaps, and space-time cubes were used to analyze hourly, daily, and seasonal trends, while hierarchical clustering and Moran's I uncovered spatial clusters and diffusion patterns.

The results reveal distinct diurnal and seasonal trends, such as higher levels of NO_2 and PM10 during morning and evening hours and elevated PM2.5 concentrations in winter. While these trends suggest links to human activities like rush hour or heating during colder months, further analysis would be required to confirm such causes. Spatially, central districts showed higher concentrations of pollutants, potentially reflecting urban density, while flow maps highlighted directional patterns in pollutant diffusion across regions. This study emphasizes the importance of integrating computational methods with human reasoning to derive actionable insights, offering guidance for policymakers and urban planners aiming to address air pollution in metropolitan areas.

1 Problem Statement

Air pollution remains one of the most pressing urban challenges, with significant implications for public health, environmental sustainability, and overall quality of life. Seoul, as a densely populated metropolitan city, experiences complex air pollution patterns influenced by factors such as traffic density, industrial activity, and seasonal changes. Addressing these patterns requires a systematic approach to analyze spatial and temporal dynamics, identify pollution hotspots, understand pollutant diffusion, and explore relationships between key pollutants.

The analysis aims to address three key questions:

- How do pollutant concentrations vary spatiotemporally in Seoul, and what temporal patterns can we detect?
- Where are the key air pollution hotspots in Seoul, and what factors explain their persistence?
- How does spatial proximity influence pollutant levels, and what are the diffusion patterns across Seoul?

The dataset used for this analysis is the 'Measurement_summary.csv', sourced from the Seoul Metropolitan Government. It contains condensed air quality measurements recorded hourly between 2017 and 2019 across 25 districts. The dataset includes six key pollutants: SO_2, NO_2, CO, O_3, PM10, and PM2.5, along with corresponding timestamps, district identifiers, and status indicators for measurement accuracy.

This dataset is well-suited for answering research questions due to its detailed temporal granularity, consistent spatial coverage, and inclusion of multiple pollutants. Its structured format allows for a robust examination of temporal trends, spatial clustering and pollutant diffusion. By combining computational analysis with visual techniques, the study aims to deliver insights that can inform evidence-based policy measures and contribute to sustainable urban air quality management.

2 State of the Art

Air pollution in urban centers like Seoul demands innovative tools to analyze spatiotemporal patterns and identify hotspots. This analysis draws on methods from six surveyed papers to explore their applicability to understanding Seoul's pollution trends, hotspots, and diffusion patterns.

Insights from the Surveyed Papers

The reviewed papers focus on urban air quality datasets, often comprising detailed measurements of pollutants like PM2.5, PM10, and NO2 collected from sensor networks. For example, Fang & Lu [3] explored temporal patterns through space-time cubes, allowing researchers to visualize how pollution evolves across both space and time. Building on this, Bach et al. [1] extended the concept, offering techniques to transform space-time data into simpler, more readable visualizations. These approaches are particularly relevant to exploring Seoul's hourly, daily, and seasonal trends.

To identify pollution hotspots, Zhou et al. [6] applied clustering techniques, such as hierarchical clustering, to reveal spatial patterns, while Gianquintieri et al. [4] used statistical tools like Moran's I to identify clusters and examine their stability over time. For diffusion patterns, Ren et al. [2] used dynamic network analysis to track how pollutants propagate between regions, and Deng et al. [5] offered interactive, graph-based visualizations that highlight relationships between districts.

Applicability to Seoul's Air Pollution Challenges

The methodologies from these papers align well with the imposed questions for analyzing air pollution in Seoul. For instance, the space-time cube models used by Fang & Lu [3] and Bach et al. [1] can effectively illustrate how pollutant concentrations change throughout the day, across seasons,

and between districts. This directly addresses the need to uncover spatiotemporal patterns and identify high-risk time windows, such as peak pollution hours.

Hotspot detection methods, like those from Zhou et al. [6] and Gianquintieri et al. [4], are equally useful. Their clustering approaches can help pinpoint areas in Seoul that consistently experience high pollution levels, such as busy traffic intersections or industrial zones. These tools can also reveal whether these hotspots are temporary or persistent throughout the year, offering valuable insights for targeted interventions.

Finally, diffusion patterns—a critical aspect of Seoul's pollution dynamics—can be explored using the network-based approaches from Ren et al. [5] and Deng et al. [2] These methods allow us to track how pollution spreads between districts, influenced by factors like wind or proximity to sources. Such insights are invaluable for designing region-specific mitigation strategies.

Limitations and Adjustments for Seoul

While these techniques are promising, their application to Seoul comes with challenges. Many of the surveyed papers assume uniform sensor coverage and consistent data quality, which may not hold true in Seoul's diverse urban landscape. Uneven distribution of monitoring stations and the city's complex topography might require adjustments to clustering and diffusion models.

Additionally, while the surveyed approaches focus heavily on spatiotemporal analysis, they often do not integrate external datasets, such as traffic density or industrial activity, which are crucial for understanding the root causes of pollution in Seoul. Incorporating such data would enhance hotspot analysis and explain why certain areas experience persistent pollution.

Another consideration is usability. Techniques like those in Deng et al. [2], which rely on interactive graph-based visualizations, are powerful but may require simplification to be accessible to policymakers or non-technical stakeholders in Seoul.

What We've Learned

From these papers, several key lessons emerge. First, space-time cubes provide a flexible and powerful way to analyze and visualize Seoul's spatiotemporal pollution patterns, offering clear insights into when and where pollution peaks occur. Second, clustering techniques and spatial autocorrelation tools are effective for identifying and analyzing pollution hotspots, especially if they can be adapted to include Seoul's unique urban factors. Finally, dynamic network approaches offer a fresh perspective on how pollutants spread across districts, revealing patterns that static maps or simpler analyses might miss.

By adapting these methods to Seoul's specific context, we can build a comprehensive framework for understanding the city's air pollution dynamics. These insights will not only deepen our understanding of the problem but also provide actionable information for decision-makers to improve air quality and public health.

3 PROPERTIES OF THE DATA

This study uses a combined dataset to analyze air quality in Seoul, created by merging three datasets provided by the Seoul Metropolitan Government's Open Data Plaza and summarized by a Kaggle contributor. The combined dataset offers a holistic view of pollutants across the city, incorporating measurements, pollutant details, and station information.

Dataset Overview

1. *Data Collection and Structure: The dataset combines:*

Measurement Info: Raw sensor data with station codes, pollutant codes, confidence scores, and hourly timestamps.

Measurement Item Info: Details about pollutants, including their names, units, value ranges, and classification criteria per regulations.

Measurement Station Info: Information about station codes, names, addresses, and geographic coordinates.

After merging, the dataset includes:

Measurement Date: Hourly averages of sensor data recorded every five minutes.

Station Code: Unique codes for Seoul's 25 monitoring stations.

Address: Exact locations of stations.

Latitude/Longitude: Geographic coordinates of stations.

Pollutants: Measurements for six key pollutants:

SO2 (Sulfur dioxide): Measured in ppm.

NO2 (Nitrogen dioxide): Measured in ppm.

O3 (Ozone): Measured in ppm.

CO (Carbon monoxide): Measured in ppm.

PM10 (Particulate matter ≤10 μm): Measured in μg/m³.

PM2.5 (Particulate matter ≤2.5 μm): Measured in μg/m³.

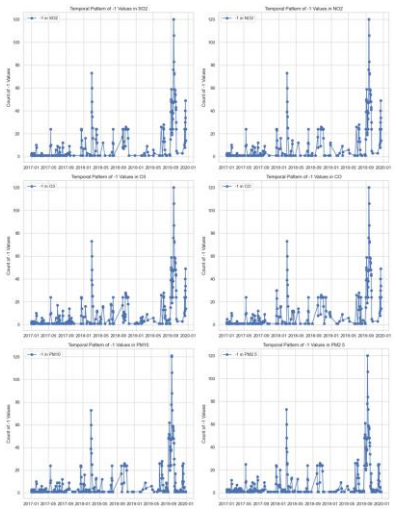
In total, the dataset has 647,511 rows and 11 columns, capturing data over time for all 25 stations.

Investigating Data Quality

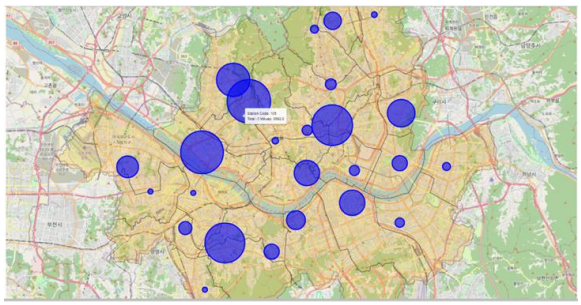
1. Missing Values: While no missing values appeared initially, further investigation revealed approximately 4,000 values of -1 in each pollutant column. Since the negative values have no meaning to our pollutant measurements we decide to treat them as missing values. To understand these:

A bar plot showed uniform counts of -1 across pollutants, suggesting dependencies between sensors.

A temporal patterns plot revealed synchronized spikes in missing values, confirming interdependence among measurement devices. (Figure below shows the temporal patterns of missing values for each pollutant.)



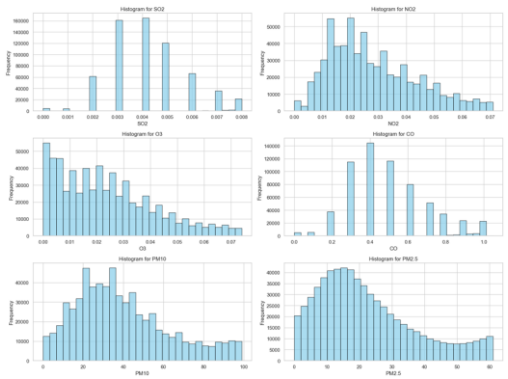
A stations map highlighted stations with higher missing value rates, but no clear geographic pattern emerged, ruling out environmental biases like proximity to rivers or altitude. (Figure below highlights this spatial distribution.)



Since missing values accounted for only ~6% of the dataset and time-series continuity is critical, we used linear interpolation to fill the gaps without introducing significant biases.

2. Outliers: Outliers were identified using box plots and the IQR method. These were consistently higher-than-expected pollutant values. Cross-referencing with confidence scores showed no link to sensor issues.

Attempts to cap outliers caused artificial spikes at the distribution limits, especially for PM10 and PM2.5. This distortion led us to replace outliers with interpolated values to maintain smooth, realistic pollutant distributions. (Figure below illustrates the histograms of pollutant distributions and the effects of outlier handling.)



Coverage of the Data

The dataset provides hourly measurements, offering high temporal resolution. Spatially, the data spans Seoul's 25 stations, evenly distributed across the city, ensuring comprehensive coverage and enabling robust spatial-temporal analysis of air quality.

Insights from Visualizations

Three key visualizations provided actionable insights:

Temporal Patterns Plot: Synchronized spikes in missing values confirmed interdependent sensor operations.

Station Map: Highlighted stations with higher missing rates but no environmental biases.

Histograms of Pollutants: Showed pollutant distributions, illustrating the effects of outlier handling and interpolation.

These insights guided decisions on how to handle missing data and outliers effectively.

Addressing Data Challenges

Missing Data: Linear interpolation ensured smooth temporal continuity.

Outliers: Replacing outliers with interpolated values avoided artificial spikes while preserving realistic distributions.

Spatial Uniformity: While some stations had higher missing rates, no spatial biases were detected, ensuring fair analysis across the city.

Conclusion

The data quality investigation laid a strong foundation for analysis. Missing values and outliers were handled thoughtfully, preserving the dataset's temporal and spatial integrity. These efforts prepare the dataset for meaningful analysis of Seoul's air quality, providing insights into pollutant patterns and supporting informed interventions.

4 ANALYSIS

4.1 Approach

In this section, the steps taken to address the research objectives are outlined. Each task focuses on a distinct aspect of air pollution analysis, combining computational techniques

and visual methods to extract meaningful insights. While the tasks are independent, they collectively contribute to a deeper understanding of pollutant dynamics in Seoul. At key stages, human reasoning plays an essential role in interpreting results, validating patterns, and guiding decisions based on computational outputs and visual representations.

Spatiotemporal Analysis

Understanding how pollutant concentrations vary across time (hourly, daily, and seasonal) and space (districts in Seoul) forms a fundamental part of the analysis. Temporal aggregation techniques will be applied to identify patterns and anomalies across different time intervals, while spatial analysis highlights disparities in pollutant levels across districts. Human reasoning is crucial for interpreting patterns in heatmaps and line plots, validating whether observed trends align with expected temporal behaviors, and identifying whether specific timeframes or anomalies require further investigation. Adjustments to temporal aggregation parameters will rely on these insights to refine analysis outcomes. Visualization tools such as line plots, heatmaps, and choropleth maps will illustrate recurring spikes and spatial patterns. Space-time cubes enable multidimensional exploration and facilitate identifying temporal and spatial trends effectively.

Hotspot Detection and Spatial Clustering

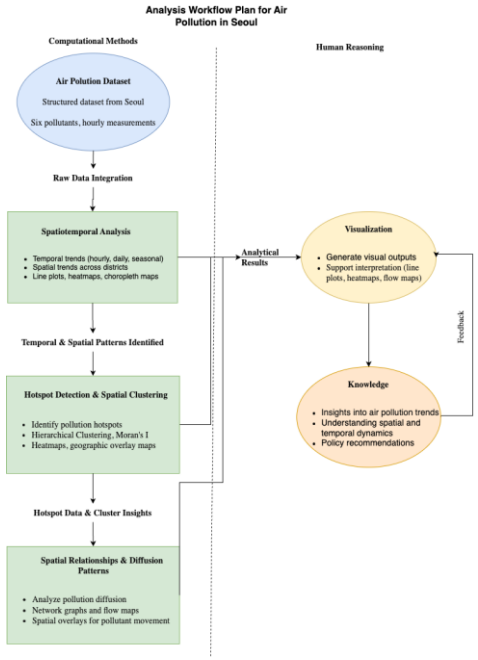
To identify areas in Seoul with consistently high pollutant levels, hierarchical clustering is used to group districts based on pollution patterns, with sensitivity analysis helping to fine-tune parameters for meaningful results. Human reasoning plays a key role in assessing whether the identified clusters align with logical spatial patterns, such as proximity to industrial zones or high traffic areas and deciding whether adjustments are necessary. Adding context through external data, like traffic density or urban land use, enhances the understanding of pollution sources and supports more informed interpretations. Visual tools like heatmaps and geographic maps clearly display the spatial distribution of clusters and their connections to external factors, making it easier to validate findings and ensure they reflect real-world conditions. This combination of computational methods and human insights creates a well-rounded approach to analyzing pollution hotspots and their potential causes.

Spatial Relationships and Diffusion Patterns

Analyzing how pollutants move between districts and how spatial relationships influence their concentrations is crucial for understanding pollutant dynamics. Pairwise differences in pollutant levels between monitoring stations are calculated to estimate flow intensity, with normalization applied to account for varying pollutant ranges. These differences form the basis of a flow network, where stations are represented as nodes and flows as directed edges, with weights reflecting the intensity and direction of movement. Human reasoning is critical for assessing whether observed patterns, such as pollutant sources or sinks, align with logical spatial behaviors. Visualizations such as directed graphs and Sankey diagrams illustrate key flow pathways and highlight areas with significant pollutant exchange. Spatial overlays on geographic maps further enhance understanding by displaying the flow network alongside pollutant concentration distributions, making it easier to identify critical pathways and areas requiring intervention. This approach combines computational precision with human interpretation to deliver actionable insights into pollutant diffusion dynamics in Seoul.

Conclusion of Approach

This approach provides a comprehensive analysis of air pollution dynamics in Seoul by integrating computational techniques with human reasoning. Computational tools efficiently process large-scale data, detect patterns, and generate precise visualizations, while human interpretation ensures logical alignment with real-world phenomena. By employing methods such as temporal aggregation, hierarchical clustering, and pollutant flow network analysis, the workflow uncovers how pollutants vary across time and space. Visualization tools, including line plots, heatmaps, and flow maps, present complex relationships in an accessible format, supporting informed decision-making and iterative refinements. This balanced integration of automation and human insight ensures that the analysis not only identifies key patterns but also contextualizes them for meaningful conclusions, delivering a clear and actionable understanding of air pollution patterns in Seoul.



4.2 Process

Temporal Analysis: Identifying Patterns Across Hours, Days, and Seasons

The analysis began with a focus on understanding the temporal distribution of pollutants across hours, days, and seasons. Using hourly average trends of normalized pollutant

concentrations (Figure 1), we observed distinct diurnal variations. NO_2 and PM_{10} peaked during morning and evening rush hours, likely linked to commuting. In contrast, O_3 concentrations peaked mid-afternoon, likely driven by sunlight-initiated photochemical reactions. These findings align with Fang et al.'s [3] application of space-time cubes for examining diurnal pollution patterns.

The sharp rise in NO_2 during rush hours suggested a strong link to vehicular emissions, supported by a weekday-weekend comparison showing significantly lower NO_2 levels on weekends. Seasonal line plots (Figure 1) revealed that $PM_{2.5}$ and SO_2 concentrations were higher in winter, potentially due to heating activities and stagnant atmospheric conditions. O_3 levels spiked in summer, consistent with enhanced photochemical activity. These observations prompted questions about whether temporal drivers were uniform across Seoul or localized to specific districts. This reasoning aligns with Bach et al.'s [1] framework for integrating temporal and environmental factors to reveal nuanced patterns.

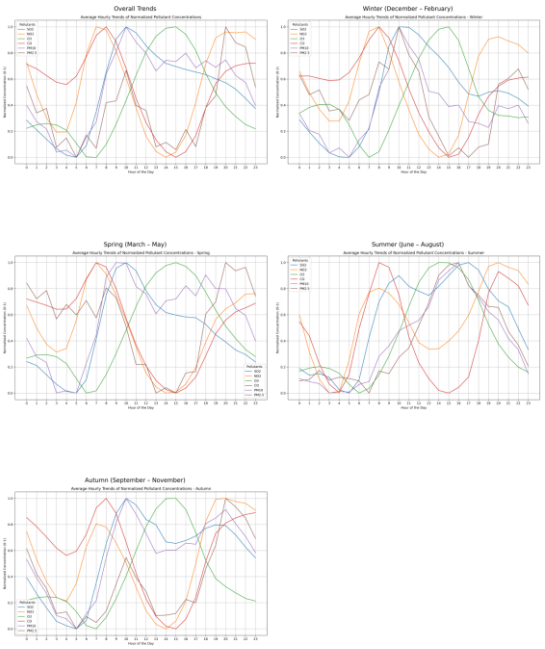


Figure 1: Overall and Seasonal Pollutant Trends in Seoul

Heatmap Analysis: Weekday and Hourly Distribution of Pollutants

To capture finer details in temporal dynamics, we created heatmaps that show pollutant distributions by hour and day, as seen in Figure 2. These visualizations highlighted weekday-weekend differences, such as reduced NO_2 and CO levels on weekends, which might be explained by less traffic. However, PM_{10} and $PM_{2.5}$ levels appeared less affected, possibly due to consistent sources like industrial activities or residential heating. These findings reflect the principles in Fang et al.'s [3] space-time analysis, which emphasized capturing both temporal and source-specific dynamics in urban pollution.

A particularly interesting trend was the early afternoon dip in SO_2 concentrations, which may be linked to emission regulations or shifts in fuel use. Another possibility is that natural dispersal processes, such as wind patterns, contribute to this dip. This aligns with the iterative data exploration approach described by Bach et al. [1], where heatmaps are used to identify temporal patterns and refine hypotheses.

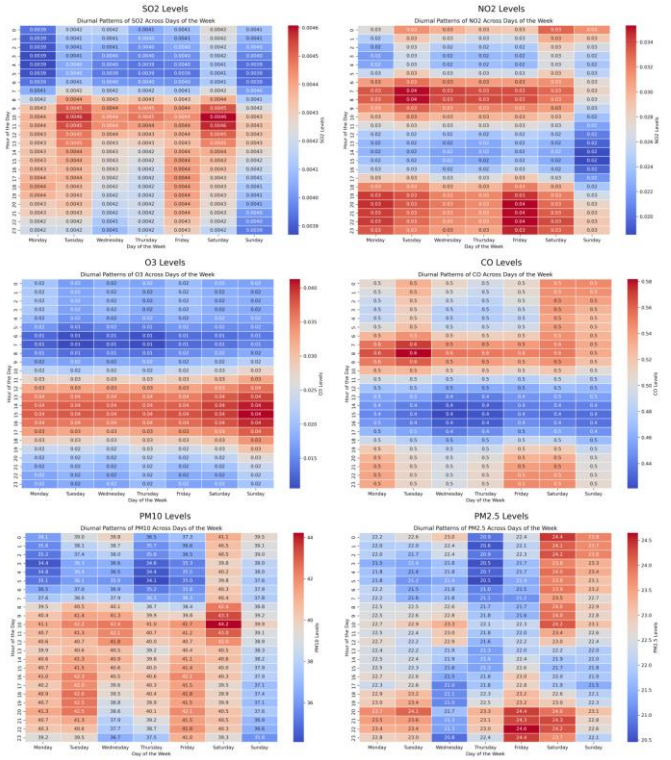


Figure 2: Diurnal Patterns of Pollutants Across Days of the Week in Seoul

Spatial Analysis: Integrating Space-Time Cubes

Next, we analyzed spatial and temporal variations using space-time cubes, shown in Figure 3. These 3D visualizations integrate pollutant data with time represented on the z-axis. Based on the geography of Seoul, we hypothesize that central areas near the river, which are likely hubs of urban density and industrial activity, may serve as hotspots for pollutants like NO_2 and PM_{10}. Meanwhile, the more rural outskirts could experience occasional spikes in pollutants such as SO_2 and CO, possibly due to specific emission sources or localized agricultural practices. This approach builds on Fang et al.'s [3] space-time cube methodology, offering a way to explore and hypothesize about pollution patterns by connecting spatial and temporal data.

To achieve meaningful results, we fine-tuned parameters like the spatial binning resolution, which was set at 0.01° to balance granularity and interpretability, and adjusted temporal aggregation to highlight seasonal trends. For instance, $PM_{2.5}$ concentrations in winter were concentrated in low-lying areas, potentially due to temperature inversions that trap pollutants close to the ground. This spatial analysis builds on Bach et al.'s [1]

recommendations for iterative parameter adjustments to uncover actionable insights from space-time data.

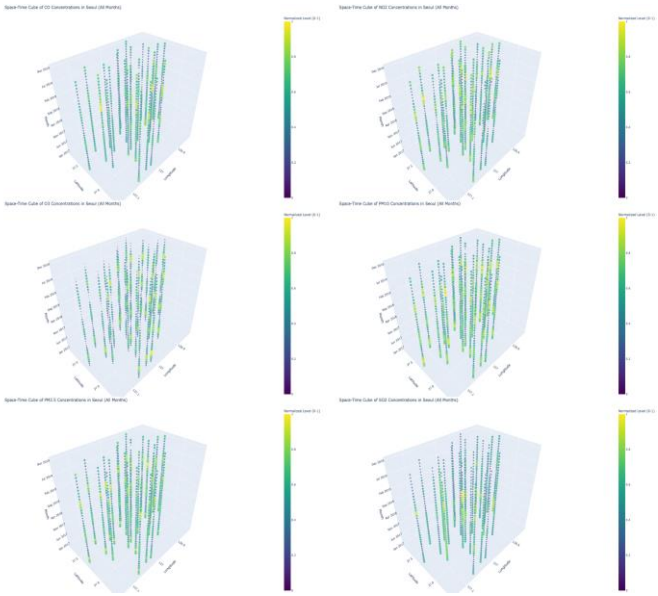


Figure 3: Space-Time Cubes of Pollutant Concentrations in Seoul

Refining Parameters and Stopping Condition

Throughout the analysis, we continuously refined methods to ensure meaningful and accurate results. Initial diurnal trends, revealed by heatmaps and line plots, led us to explore seasonal and spatial patterns. Space-time cubes further integrated these dimensions, offering a unified view of pollutant evolution across time and space.

By normalizing pollutant concentrations and testing different visualization perspectives, we ensured consistency in spatial patterns. Additional iterations focused on validating seasonal trends and identifying areas where anomalies persisted, adding robustness to the findings. This iterative process aligns with both Fang et al. [3] and Bach et al. [1], who emphasize the importance of validating results against observed trends and revisiting assumptions.

We concluded the analysis when patterns from different visualizations aligned, providing a comprehensive understanding of temporal and spatial drivers. This rigorous approach ensured consistency and reliability in identifying pollutant behaviors across Seoul.

Spatial and Temporal Pollutant Patterns

We began the analysis by standardizing the pollutant concentration data. This ensured uniformity and compatibility with time-dependent operations such as grouping. Standardization was particularly advantageous given the near-normal distribution observed during preprocessing. This choice aimed to prevent bias in clustering or spatial analyses caused by the varying magnitudes of pollutants, aligning with best practices in data science.

To introduce temporal granularity, we aggregated the data into daily, monthly, and seasonal averages. This transformation was essential for detecting periodic trends and allowed for simultaneous exploration of patterns across the dataset, offering insights into both short-term and long-term pollutant behaviors.

Dimensionality Reduction and Clustering

To uncover potential clusters, we applied Multidimensional Scaling (MDS), which reduced the high-dimensional pollutant data into a more interpretable 2D space. While early projections showed no clear separation, we opted for hierarchical clustering with Ward's method to delve deeper into the data structure. This method, which minimizes within-cluster variance, revealed hierarchical relationships through a dendrogram. Interpreting the dendrogram, we identified large vertical gaps as indicators of meaningful separation, refining our choice of cluster numbers.

After redefining the clusters, we visualized them again using MDS, observing more distinct cluster formations. To link these clusters to their corresponding station locations, we plotted them spatially. Despite this improvement, overlapping points persisted, making it challenging to determine specific clusters. At this stage, human judgment played a critical role in deciding to incorporate spatial autocorrelation measures, such as Global and Local Moran's I.

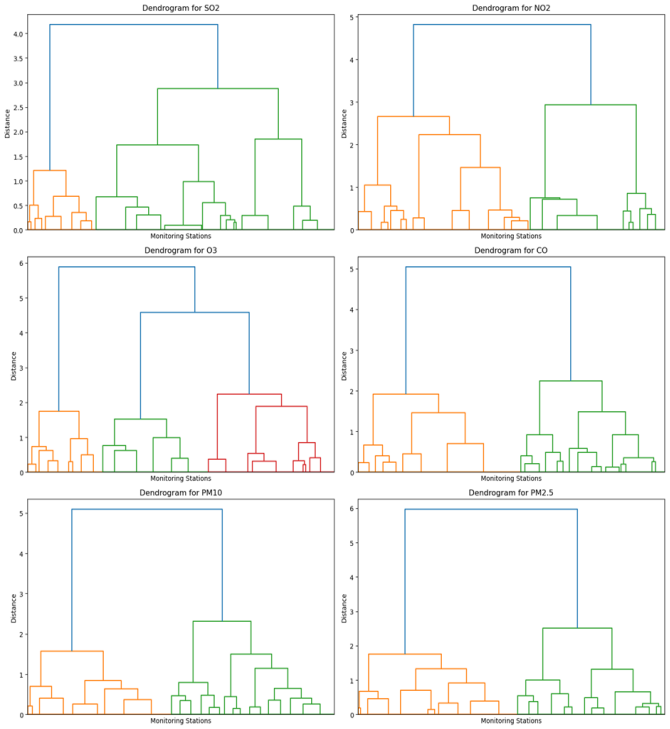


Figure 4: Dendrograms of all different 6 pollutants

Spatial Autocorrelation and Refinement

Using distance-based weights, we applied Moran's I to evaluate the spatial similarity of pollutant levels across locations. To ensure robust results, we validated findings with p-values, identifying significant hotspots and coldspots. We iteratively adjusted the distance thresholds to derive meaningful spatial patterns. Focusing on one pollutant at a time, we plotted refined clusters that clearly separated stations

into three categories: hotspots, coldspots, and non-significant areas. This step provided a breakthrough in mapping pollutant clusters to Seoul's urban layout. Notably, coldspots aligned with green areas, while hotspots matched industrial zones, a result that added interpretive depth to the findings.

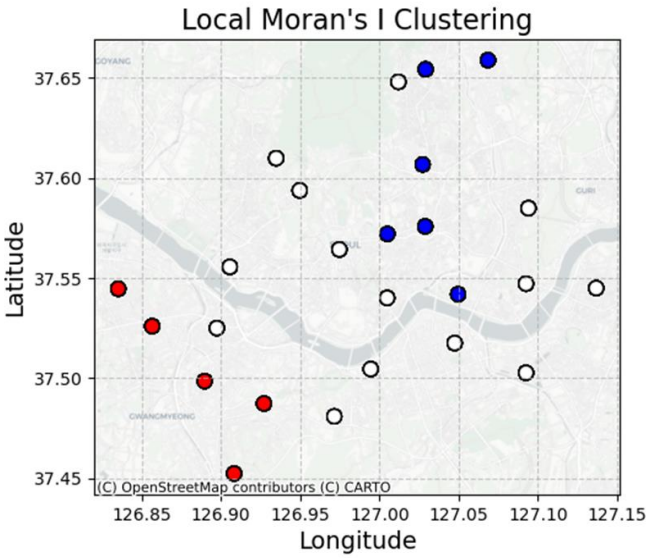


Figure 5: Hotspots and Coldspots final clustering using Local Moran's I

Voronoi Diagrams and Temporal Trends

To model spatial proximity, we used Voronoi diagrams. These polygons, integrated into a geodataframe, provided insights into how stations interacted with their surroundings. Comparing Voronoi-based clusters with Moran's I results revealed overlapping patterns, though further refinement was required for clearer interpretation.

Next, we decomposed daily pollutant data using LOESS (Locally Weighted Scatterplot Smoothing). This technique removed noise and highlighted temporal trends. Through iterative tuning of the fraction parameter, finalized at frac=0.105, we identified seasonal peaks for each pollutant. Scatterplots revealed that while most pollutants exhibited aligned fluctuations, O_3 had an inverse trend, marking a significant discovery warranting further investigation.

Using Global Moran's I, we explored spatial clustering over specific periods, identifying persistent patterns. Visualizing these values provided insights into the evolution of spatial clustering, complemented by bar plots and scatterplots showing hotspot consistency over time. These analyses offered actionable insights for authorities to target high-severity clusters.

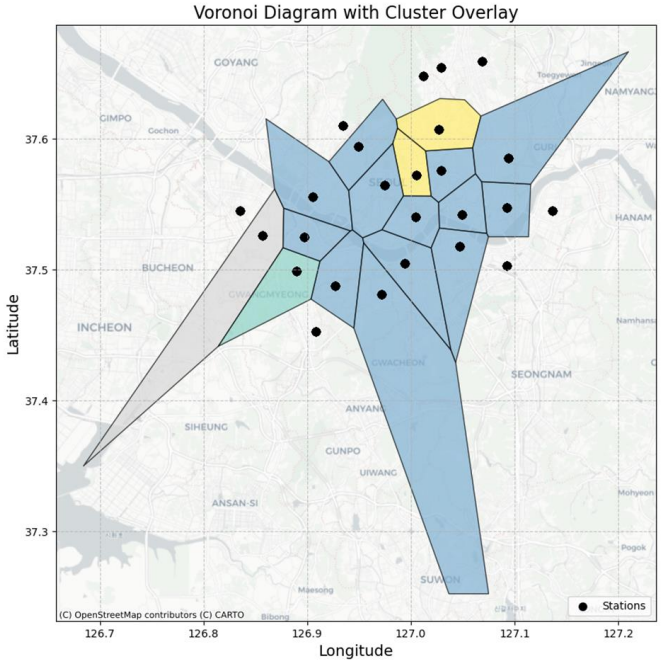


Figure 6: Voronoi diagram with cluster overlay

Pollutant Flow Network Analysis

For the third question, we leveraged the monthly aggregated data from the previous analyses to simplify and enhance interpretability. We computed concentration differences between station pairs for each pollutant, defining these as pollutant flows. Positive flows indicated potential pollutant sources, while negative flows suggested recipients. To ensure consistency, we normalized these values and stored them in a structured dataframe.

Threshold Filtering and Network Graph Construction

To manage complexity, we introduced a flow threshold, allowing us to filter significant flows dynamically. Adjusting this threshold facilitated the creation of a more interpretable network graph. Nodes were aligned with station coordinates, and edges represented pollutant flows, with weights and pollutant types added as attributes. By incorporating edge weights, we enabled advanced analyses, such as identifying hubs or visualizing flow intensities.

Initial visualizations included all pollutants, but we quickly realized that separating them would provide sharper insights. Using high thresholds, we produced pollutant-specific graphs, replacing intensity color bars with pollutant-specific indicators. This refinement allowed for pollutant-specific analyses, uncovering distinct patterns.

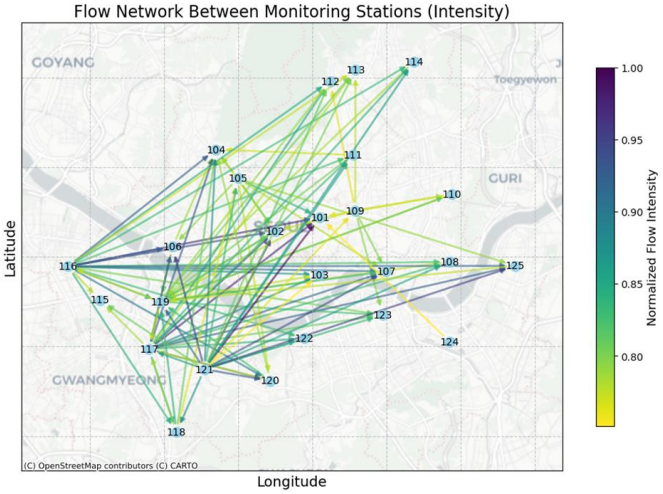


Figure 7: Flow Network emphasizing on the intensity of flow

Temporal Dynamics and Sankey Diagrams

To examine temporal dynamics, we created static network graphs for each month over a three-year span. Sequentially visualizing these graphs revealed changes in pollutant flow directions and highlighted months with increased significant flows. This approach bridged static and dynamic analyses, offering a comprehensive view of pollution trends.

To enhance visual interpretation, we constructed a Sankey diagram using Plotly. Interactive feedback guided the refinement of attributes such as padding and thickness. This visualization intuitively traced pollutant movements, identifying major source and target stations. The Sankey diagram effectively summarized complex pollutant flow patterns, highlighting contributors and recipients in an easily interpretable format.

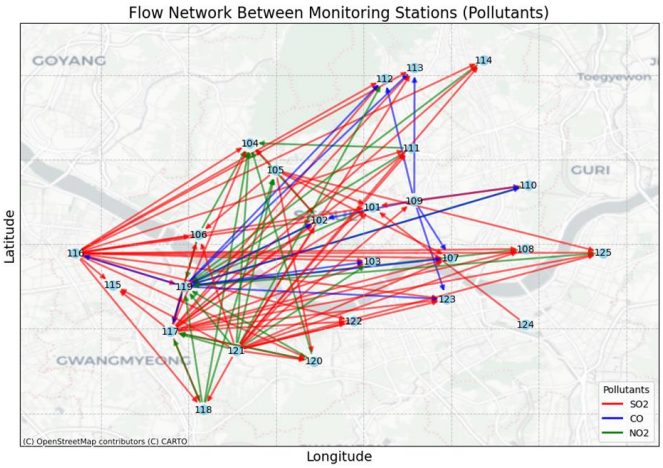


Figure 8: Flow Network emphasizing on the pollutants

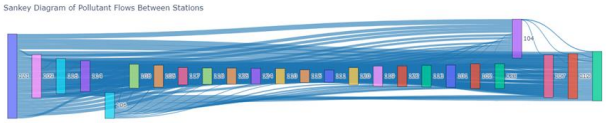


Figure 9: Sankey interactive diagram categorizing stations (high to low pollution sources) and showing flows.

Key Insights and Stopping Conditions

Throughout the analysis, we used iterative feedback loops between visual and computational methods to refine our approaches. Decisions, such as adjusting clustering parameters or flow thresholds, were informed by both domain knowledge and insights gained from intermediate results. The stopping condition for each analysis phase was determined by achieving clear, interpretable results that aligned with the study's objectives.

For clustering, we stopped after hierarchical clustering and spatial autocorrelation measures consistently produced meaningful clusters that mapped logically to Seoul's urban layout. Similarly, pollutant flow analyses concluded when visualizations, such as Sankey diagrams, effectively highlighted critical patterns without overwhelming complexity.

By combining computational rigor with human interpretation at each step, we ensured that the findings were robust, actionable, and aligned with the research questions.

4.3 Results

Aggregated Pollution Levels Across the Years

When exploring the data on aggregated pollution levels, we noticed clear differences in where each pollutant tends to concentrate. These distinctions are well illustrated in the six pollutant maps we generated during the spatiotemporal analysis. While we initially wanted to overlay the map of Seoul onto a 3D space-time cube for a more dynamic view, the visual tools we had available made this challenging. Instead, we opted for 2D maps, which still effectively showcase the results and highlight the spatial trends.

Figure 10: 2D overlayed heatmaps

Temporal Trends of Pollutants

To dig deeper into temporal patterns, we used LOESS to identify trends for each pollutant over time. Most pollutants showed a similar pattern, but O_3 stood out. It followed an opposite trend or perhaps a shifted one compared to the others. This might be because O_3 forms higher up in the atmosphere, so the measurements are delayed. This observation is intriguing and could benefit from further exploration by experts in atmospheric science who can investigate these dynamics in more detail.

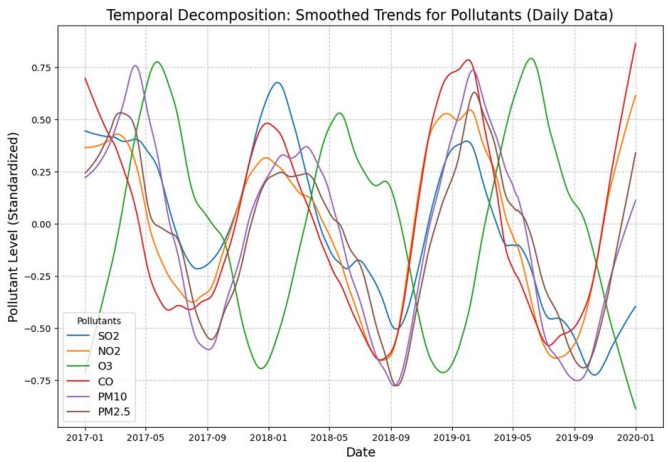


Figure 11: Smoothened scatterplot showing temporal trends

Network Graph and Spatial Dynamics

To understand how pollution flows between locations, we built a network graph with a strict flow threshold to make the results clearer. The graph revealed a strong pattern of outgoing arrows moving from the left toward the upper-right, suggesting hotspot stations where pollution originates and coldspot stations where it tends to collect. These findings align with the clusters we saw using local Moran's I analysis and even hint at wind patterns in the city. Unfortunately, we didn't have access to wind and weather data, which would have been helpful for validating and expanding on these results.

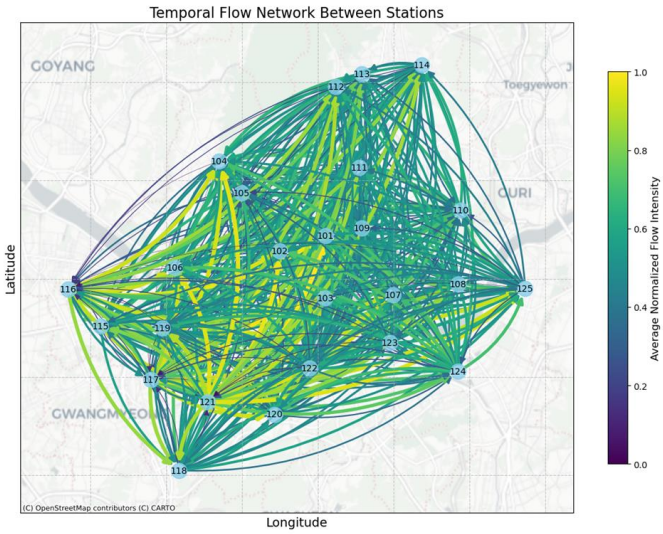


Figure 12: Temporal Flow Network Graph of Seoul Stations

Implications and Limitations

These findings provide valuable insights into how pollution behaves across space and time in the city. Policymakers could use this information to prioritize areas for intervention, like reducing pollution in hotspots. The lack of wind and weather data, as well as the limitations in our visualization tools, means there's more work to be done.

5 CRITICAL REFLECTION

This analysis provided valuable insights into Seoul's air pollution dynamics, but the process revealed several challenges and areas for improvement. Visual representations played a critical role, uncovering patterns that computational methods alone could not. However, limitations in data availability, techniques, and tools influenced the scope of our findings, highlighting opportunities for refinement.

The temporal analysis captured hourly, daily, and seasonal patterns, but it relied heavily on human reasoning to interpret results and refine hypotheses. For example, diurnal spikes in NO_2 during rush hours were evident in heatmaps, but their correlation with traffic density required external datasets. Integrating traffic density data could provide direct evidence for these correlations, reducing the reliance on assumptions. Similarly, winter peaks in PM10 and SO_2 appeared linked to heating activities. Including energy usage or emissions data related to heating could validate these observations, making the analysis more robust. Employing machine learning models trained on such enriched datasets could also deepen insights and identify hidden patterns.

Spatial analysis added complexity by revealing pollutant distribution across Seoul using space-time cubes, which highlighted potential hotspots near urban centers and localized spikes in peripheral areas. However, the iterative adjustments required for parameters like spatial binning and temporal aggregation were time-intensive and subjective. Optimization algorithms for parameter selection based on statistical metrics could streamline this process. Dynamic spatial-temporal modeling frameworks may further enhance

the precision and scalability of analyses, offering a more systematic approach to interpreting spatial patterns.

Cluster analysis using hierarchical methods and Moran's I identified pollution hotspots but highlighted challenges with overlapping data points and unclear cluster boundaries. These issues could be mitigated by incorporating auxiliary datasets, such as land use maps or population density metrics, to contextualize pollution sources. Advanced clustering techniques, such as density-based spatial clustering (e.g., DBSCAN) or fuzzy clustering, could provide clearer delineations. Voronoi diagrams offered a useful spatial perspective but oversimplified interactions between monitoring stations and their surroundings. Combining these diagrams with granular geospatial data or integrating them into a broader geostatistical framework could improve the reliability of insights.

The pollutant flow network offered a novel perspective on diffusion patterns but came with technical limitations. Simplifying the data to highlight meaningful flows risked overlooking subtle patterns. Including meteorological data, such as wind direction and speed, could greatly enhance the accuracy of flow direction predictions and uncover underlying causes. Advanced network modeling techniques, such as weighted or stochastic flow simulations, could capture lower-intensity pathways without compromising clarity. Interactive Sankey diagrams that allow threshold adjustments could also provide more nuanced insights while maintaining interpretability.

Reflecting on the overall process, the importance of balancing computational methods with human insight became evident. While computational tools efficiently processed large-scale data, human reasoning was essential for interpreting results and ensuring their alignment with real-world phenomena. Visual tools like heatmaps, space-time cubes, and Sankey diagrams effectively bridged data and conclusions but also highlighted the need for more intuitive, interactive tools to support iterative exploration. Developing platforms that integrate domain-specific datasets and facilitate real-time visualization adjustments could significantly enhance analytical workflows.

Future analyses could benefit from integrating additional datasets, such as traffic density, meteorological conditions, or industrial activity records, to provide richer context for observed patterns. Enhanced software functionality, such as automated clustering validation or dynamic spatial modeling, could streamline workflows and reduce manual adjustments. Leveraging cloud-based analytics platforms for real-time data integration and analysis would address data gaps and improve scalability. Additionally, creating accessible visualizations tailored for policymakers and non-technical stakeholders could help translate findings into actionable strategies for air quality management.

In conclusion, this analysis successfully uncovered spatiotemporal pollution patterns and identified key hotspots and diffusion pathways in Seoul. However, the study's limitations underscore the need for comprehensive datasets, methodological adaptability, and a synergy between computational precision and human judgment. By addressing these challenges and adopting the proposed solutions, future studies can build on this foundation to deliver more impactful insights into urban air quality dynamics

Table of word counts

Problem statement	248
State of the art	653
Properties of the data	618
Analysis: Approach	630
Analysis: Process	1573
Analysis: Results	329
Critical reflection	646

REFERENCES

[1] Bach, B., Dragicevic, P., Archambault, D., Hurter, C. and Carpendale, S. (2016). A Descriptive Framework for Temporal Data Visualizations Based on Generalized Space-Time Cubes. Computer Graphics Forum, 36(6), pp.36–61. doi:https://doi.org/10.1111/cgf.12804.

[2] Deng, Z., Weng, D., Chen, J., Liu, R., Wang, Z., Bao, J., Zheng, Y. and Wu, Y. (2019). AirVis: Visual Analytics of Air Pollution Propagation. IEEE Transactions on Visualization and Computer Graphics, 26(1), pp.1–1. doi:https://doi.org/10.1109/tvcg.2019.2934670.

[3] Fang, T.B. and Lu, Y. (2011). Constructing a Near Real-time Space-time Cube to Depict Urban Ambient Air Pollution Scenario. Transactions in GIS, 15(5), pp.635–649. doi:https://doi.org/10.1111/j.1467-9671.2011.01283.x.

[4] Gianquintieri, L., Mahakalkar, A.U. and Caiani, E.G. (2024). Exploring Spatial–Temporal Patterns of Air Pollution Concentration and Their Relationship with Land Use. Atmosphere, 15(6), p.699. doi:https://doi.org/10.3390/atmos15060699.

[5] Ren, K., Wu, Y., Zhang, H., Fu, J., Qu, D. and Lin, X. (2020). Visual Analytics of Air Pollution Propagation through Dynamic Network Analysis. IEEE Access, 8, pp.1–1. doi:https://doi.org/10.1109/access.2020.3036354.

[6] Zhou, Z., Ye, Z., Liu, Y., Liu, F., Tao, Y. and Su, W. (2017). Visual Analytics for Spatial Clusters of Air-Quality Data. IEEE Computer Graphics and Applications, 37(5), pp.98–105. doi:https://doi.org/10.1109/mcg.2017.3621228.

[7] draw.io (2024). Flowchart Maker & Online Diagram Software. [online] app.diagrams.net. Available at: https://app.diagrams.net/.

[8] Seoul.go.kr. (2019). 서울시 대기오염 측정정보. [online] Available at: https://data.seoul.go.kr/dataList/OA-15526/S/1/datasetView.do [Accessed 11 Jan. 2025].

[9] www.kaggle.com. (n.d.). Air Pollution in Seoul. [online] Available at: https://www.kaggle.com/datasets/bappekim/air-pollution-in-seoul.

[10] Github.io. (2024). Using GeoJson — Folium 0.18.0 documentation. [online] Available at: https://python-visualization.github.io/folium/latest/user_guide/geojson/geojson.html.

[11] geopandas.org. (n.d.). geopandas.GeoDataFrame — GeoPandas 0.14.4+0.g60c9773.dirty documentation. [online] Available at: https://geopandas.org/en/stable/docs/reference/api/geopandas.GeoDataFrame.html.

[12] docs.scipy.org. (n.d.). scipy.spatial.Voronoi — SciPy v1.13.1 Manual. [online] Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.Voronoi.html.

[13] docs.scipy.org. (n.d.). fcluster — SciPy v1.14.0 Manual. [online] Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html.

[14] www.statsmodels.org. (n.d.). statsmodels.nonparametric.smoothers_lowess.lowess - statsmodels 0.15.0 (+44). [online] Available at:

https://www.statsmodels.org/dev/generated/statsmodels.nonparametric.smoothers_lowess.lowess.html.
[15] scikit-learn. (2025). MDS. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html [Accessed 12 Jan. 2025].
[16] Readthedocs.io. (2019). Introduction guide to contextily — contextily 1.6.3.dev2+geafeb3c.d20241118 documentation. [online] Available at: https://contextily.readthedocs.io/en/latest/intro_guide.html [Accessed 12 Jan. 2025].
[17] Github.io. (2019). plotly.graph_objects.Sankey — 5.24.1 documentation. [online] Available at: https://plotly.github.io/plotly.py-docs/generated/plotly.graph_objects.Sankey.html [Accessed 12 Jan. 2025].
[18] Pysal.org. (2018). esda.Moran_Local — esda v2.6.0 Manual. [online] Available at: https://pysal.org/esda/generated/esda.Moran_Local.html [Accessed 12 Jan. 2025].
[19] Pillow (PIL Fork). (n.d.). ImageDraw Module. [online] Available at: https://pillow.readthedocs.io/en/stable/reference/ImageDraw.html.