Spatial Data Analytics

Mallikarjun Chakoti
IMT2022116
IIIT - Bangalore
Bangalore , India
Mallikarjun.Chakoti@iiitb.ac.in

Ullas G
IMT2022125
IIIT - Bangalore
Bangalore , India
Ullas.G@iiitb.ac.in

I. Introduction

This report contains the pre-processing and exploratory data analysis of the Egypt Internet Speed and Performance Q4 2023 Dataset [1]. .

II. Dataset

A. Overview

The dataset contains network performance measurements collected from various locations across multiple cities in Egypt. It comprises 17,871 observations (rows) with 10 features (columns) recording download/upload speeds, latency, geographic coordinates, and other metadata.

B. Data Description

The dataset includes the following columns:

Column Name	Data Type	Description
avg_d_kbps	Numeric	Average download speed in kbps
avg_u_kbps	Numeric	Average upload speed in kbps
avg_lat_ms	Numeric	Network latency in milliseconds
tests	Integer	Number of network tests conducted
devices	Integer	Number of unique devices tested
centroid_lat	Numeric	Latitude (WGS84)
centroid_lon	Numeric	Longitude (WGS84)
city	String	City where tests were conducted
avg_d_speed_mbps	Numeric	Download speed in Mbps
avg_u_speed_mbps	Numeric	Upload speed in Mbps

Table I: Dataset Variables Description

C. Key Characteristics

- **Spatial Coverage**: Measurements span multiple cities with precise geographic coordinates
- **Network Metrics**: Comprehensive recording of download/upload speeds and latency
- **Test Metadata**: Information about the number of tests and devices used
- **Derived Values**: Includes both raw (kbps) and converted (Mbps) speed measurements

III. Spatial analytics

A. Initial analysis

1) Data Preparation and Cleaning: The dataset was initially examined for null or missing values. After thorough inspection, it was confirmed that the dataset was clean, containing neither null entries nor missing values. This allowed for direct progression to the visualization and analysis stages without the need for data imputation or row elimination.
Before delving into the spatial statistics and data mining aspects of our project, we performed an initial exploratory analysis to gain a basic understanding of the dataset. We began by generating a variety of plots to visualize the distributions and relationships among different features. By examining these plots, we were able to familiarize ourselves with the nature of the features we were working with.
Following this verification, we proceeded to visualize the **Spatial Distribution of Network Performance Measurements in Egypt**. This visualization aimed to provide a preliminary understanding of the geographical spread and density of the data points.

2) Spatial Distribution Visualization: From Figure 1, it is evident that most network performance measurements are concentrated in the Nile Delta and Greater Cairo regions. These areas are densely populated and likely have more robust network infrastructure, explaining the higher frequency of tests. Conversely, sparse measurements are observed in the western desert and southern parts of the country, which could be attributed to lower population density or limited network infrastructure.

This spatial distribution plot was selected as the first visualization to understand the geographic coverage and potential biases in the dataset. It helps identify regions with dense or sparse data availability, thus guiding subsequent detailed analyses of network performance. By starting with this plot, we establish the geographical context necessary for more nuanced interpretations of performance metrics like download speed, upload speed, and latency.

3) Distribution Plots of upload and download speed:
Download Speed Distribution:
This plot in Figure 2 shows the distribution of average download speeds (in megabits per second) across different geographic observations. It helps assess the overall performance of downstream data transfer in the network.

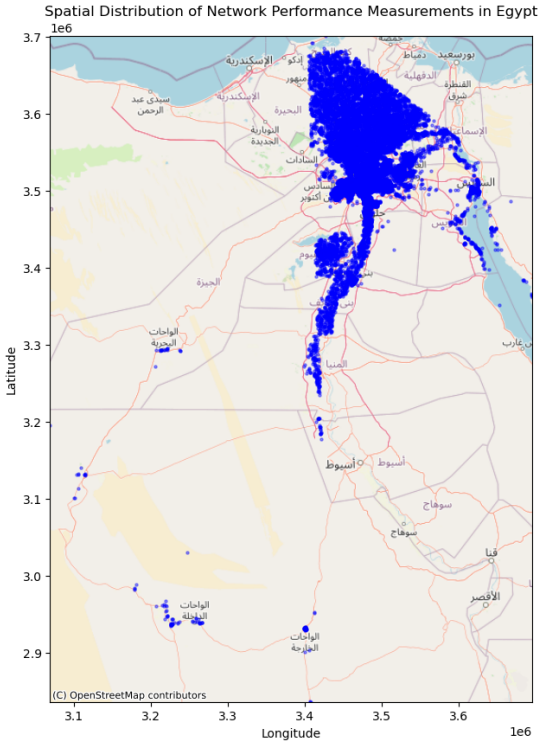

Fig. 1: Spatial Distribution of Network Performance Measurements in Egypt

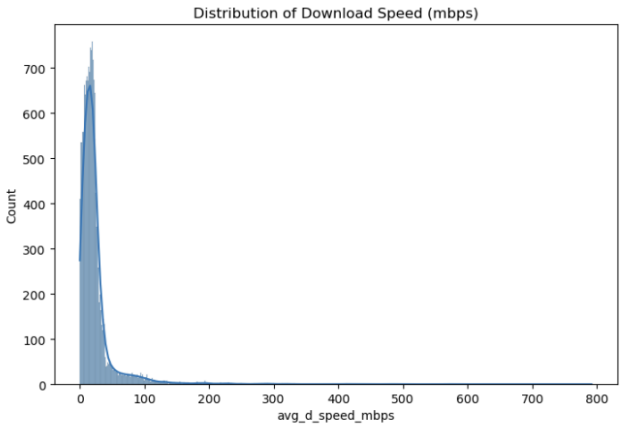

Fig. 2: Distribution of average download speeds (avg_d_speed_mbps) across measurement locations.

Inference:

- The distribution of download speeds is highly right-skewed, with the majority of values concentrated below 50 Mbps.
- A sharp peak is observed around 20–30 Mbps, indicating that most users experience speeds in this range.
- Very few observations exist above 100 Mbps, suggesting high-speed internet is rare.
- The long tail extending towards 800 Mbps represents outliers or select urban areas with advanced network infrastructure.
- This highlights a digital inequality where most regions still operate at moderate speeds while only a few access ultra-high-speed networks.

Upload Speed Distribution:

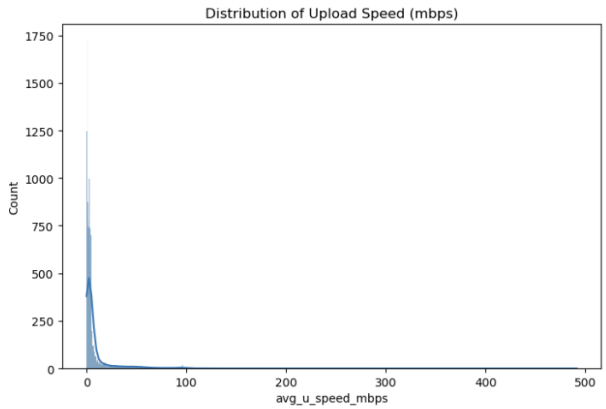

Fig. 3: Distribution of average upload speeds (avg_u_speed_mbps) across measurement locations.

This plot in Figure 3 shows the distribution of average upload speeds (in megabits per second) across different geographic observations. It helps assess the overall performance of upstream data transfer in the network.

Inference:

- The distribution is heavily right-skewed, with most upload speeds clustered below 20 Mbps.
- A sharp peak between 2–10 Mbps indicates that the majority of users experience low to moderate upload performance.
- Very few users achieve upload speeds above 50 Mbps, suggesting limited upstream capacity in most network deployments.
- The long tail up to 500 Mbps likely represents isolated high-performance areas with fiber or 5G connectivity.

4) Download Speed Distribution by City: This box plot in Figure 4 visualizes the distribution of average download speeds, measured in megabits per second (mbps), across a variety of cities. Each box represents the range and central tendency of download speeds observed within a specific city, allowing for a comparison of network performance across different geographical locations.

Inference The box plot reveals significant variations in download speed distributions among the listed cities. Some cities exhibit relatively consistent and often higher average download speeds, indicated by shorter boxes and higher median lines (the horizontal line within each box). In contrast, other cities show wider distributions, suggesting greater variability in download speeds experienced by users within those areas. Several cities also display outliers (represented by diamond shapes), indicating instances of unusually high or low download speeds that deviate significantly from the typical range for those locations.

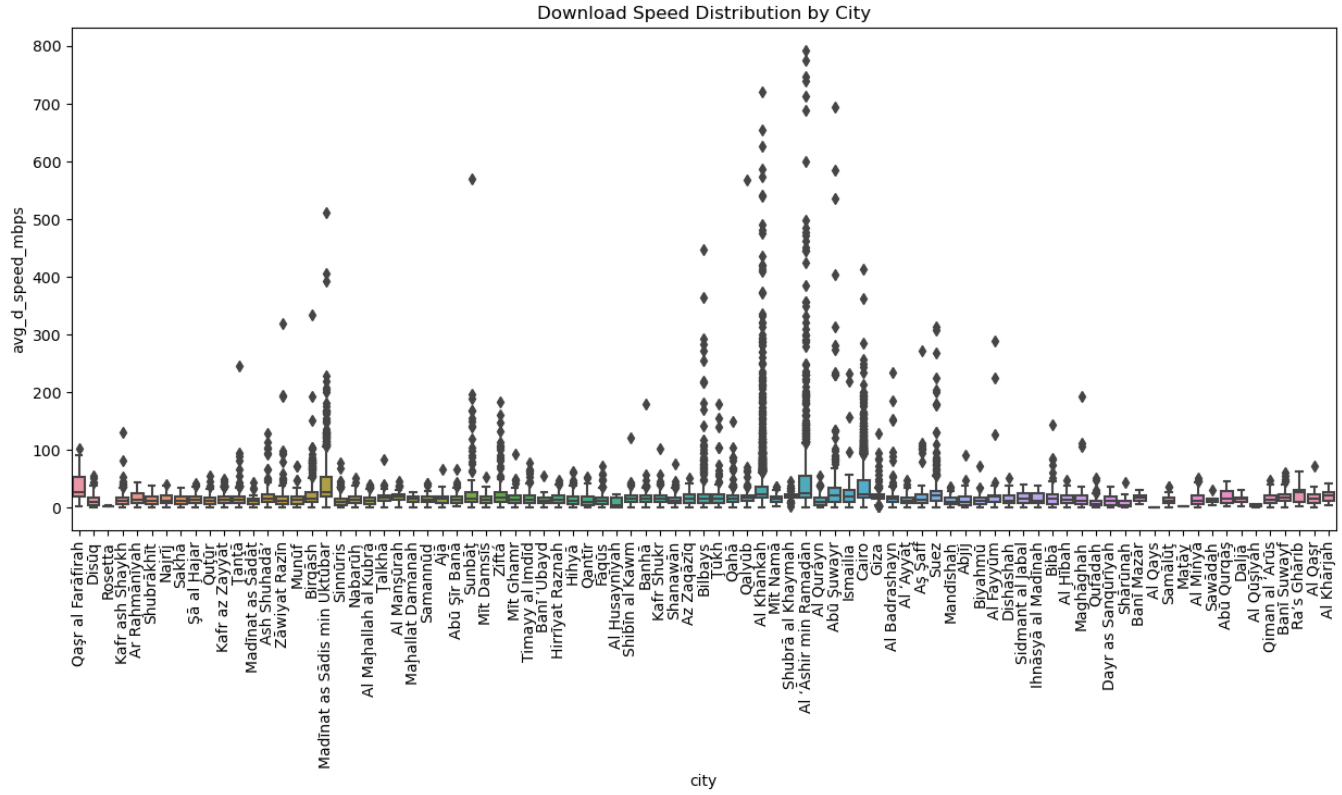

Fig. 4: Distribution of Average Download Speeds (mbps) Across Cities

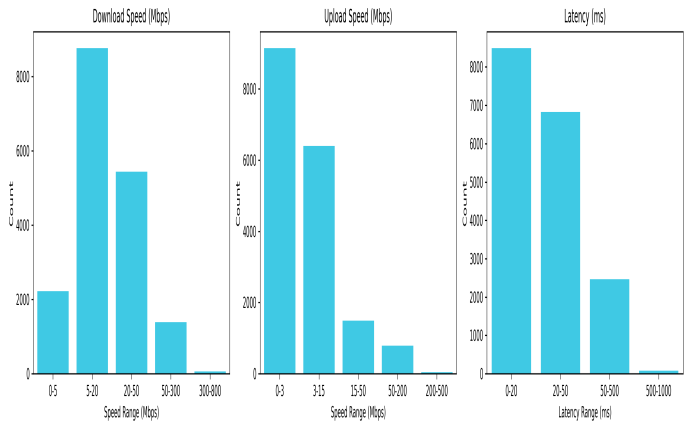

Fig. 5: Distribution of Download Speeds (Mbps) and Latency (ms) across network observations.

5) Speed latency Count Plots: The count plots in Figure 5 show how our data is spread across different ranges. The left plot displays the distribution of download speeds. From it, we can see that most data points fall in the 5–20 Mbps range, followed by the 20–50 Mbps range. Very few points have download speeds greater than 300 Mbps. The middle plot shows a similar distribution for upload speeds, where over 8000 points fall into the 0–3 Mbps range, which is quite low. Lastly, the right plot shows how latency (in milliseconds) is spread. Most values are in the 0–20 ms range, which is very good, though a fair number also lie between 50–500 ms.

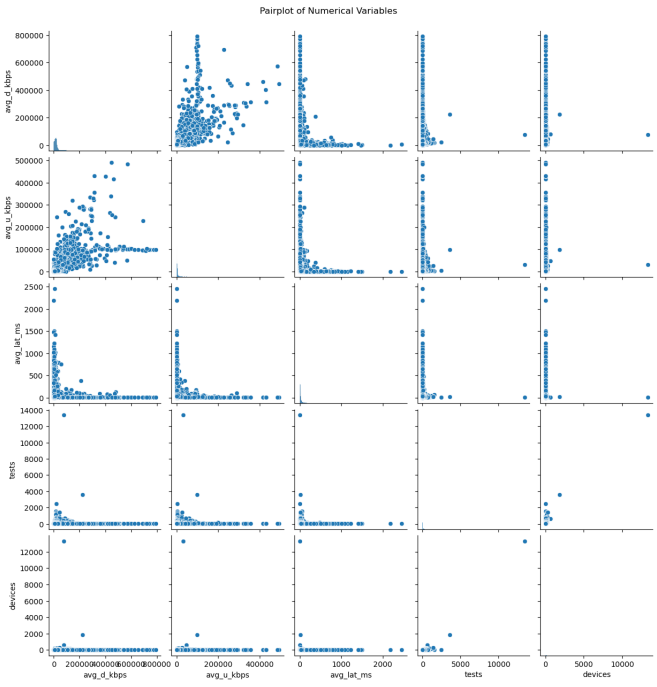

Fig. 6: Pairplot showing relationships among key numerical variables, highlighting trends and potential outliers in network performance data.

6) Pairwise Relationship Analysis of Numerical Variables: **Inference:**

- The bubble plot (Upload Speed vs Download Speed) shows a generally linear rising trend — though not perfectly linear, it indicates that an increase in upload speed is typically associated with an increase in download speed.
- A strong positive correlation is observed between `avg_d_kbps` (average download speed) and `avg_u_kbps` (average upload speed), indicating that regions with higher download speeds generally also have higher upload speeds.
- Both `avg_d_kbps` and `avg_u_kbps` show a negative correlation with `avg_lat_ms` (average latency), implying that better speed is often associated with lower latency—a desirable network characteristic.
- The scatter plots reveal clustering of data points at lower value ranges across most variables, hinting at skewed distributions and the presence of outliers.
- The variables `tests` and `devices` exhibit some extreme values (outliers), suggesting variability in user/device density or testing frequency in certain regions.
- Overall, the pairplot highlights key relationships and potential collinearity among the features, which can inform future modeling and feature selection strategies.

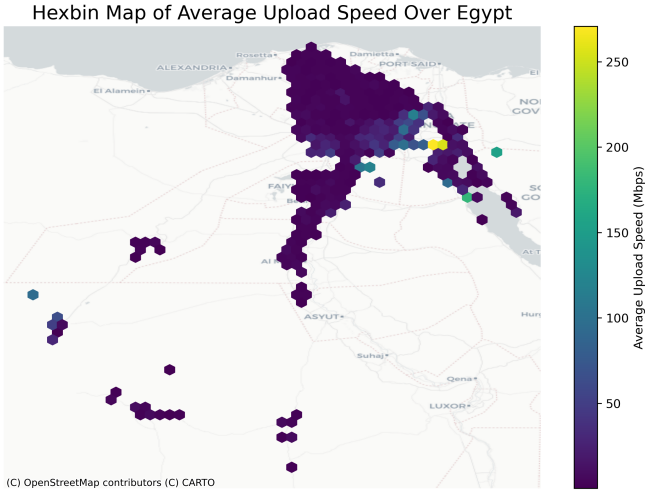

Fig. 7: Hexbin map showing the spatial distribution of upload speeds

7) Hex-Bin Plot: The Figure 7 shows the average upload speed at different locations across Egypt, based on our dataset of network measurements. To visualize the data, we used a hexbin map, which divides the map into hexagonal grid cells. Each hexagon's color represents the average upload speed (in Mbps) of all data points within that area, as shown by the color bar on the right.

Initially, we attempted to use a scatter plot with the same color scale to show upload speeds at each data point. However, because there were so many points, the scatter plot was too cluttered and difficult to interpret. To address this, we switched to a hexbin map with a grid size of 50. This approach groups nearby points together and shows the average value for each hex cell, making the map much clearer and easier to understand. The hexbin map provides a good estimate of upload speed patterns across Egypt while reducing visual clutter and highlighting regional differences more effectively. The map reveals that upload speeds are generally higher in a small section of urban and densely populated regions, such as Cairo and the Nile Delta, as indicated by the lighter colors. other regions have low upload speeds in general.

B. Spatial autocorrelation

Spatial Autocorrelation refers to the statistical relationship between the values of a variable and their spatial locations. In simpler terms, it measures the degree to which similar values occur near each other in space.

Spatial autocorrelation helps identify patterns of clustering or dispersion across geographic regions. A positive spatial autocorrelation indicates that similar values (e.g., high or low) are located near each other, while a negative value suggests that dissimilar values are neighbors. When spatial autocorrelation is absent (value near zero), it implies a random spatial distribution.

This concept is fundamental in geographical data analysis, as it allows researchers to assess the presence of spatial structures and make more informed decisions regarding spatial processes and modeling strategies.

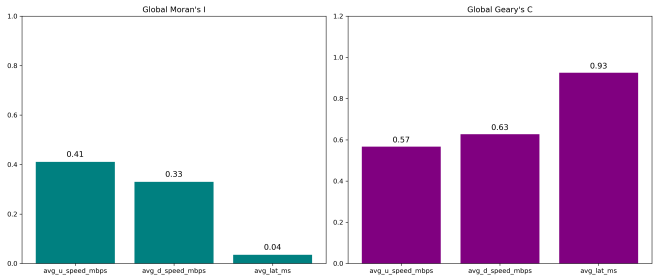

Fig. 8: Global Moran's I and Geary's C statistics indicating spatial autocorrelation in download speeds.

1) Global morons I and Geary's C : **Global Moran's I and Geary's C:** These are measures of spatial autocorrelation used to assess whether similar values are clustered (positive autocorrelation) or dispersed/random (negative or no autocorrelation).

- **Moran's I** ranges from -1 (perfect dispersion) to $+1$ (perfect clustering). Values closer to $+1$ indicate strong positive spatial autocorrelation.
- **Geary's C** ranges from 0 to 2. Values closer to 0 suggest strong clustering, while values near 1 imply spatial randomness.

Key Findings from the Plot in Figure 8:

- **Upload Speed (`avg_u_speed_mbps`):**

- Moran's I = 0.41 ⇒ Strong positive spatial autocorrelation (regions with fast/slow speeds tend to cluster).
- Geary's C = 0.57 ⇒ Indicates clustering (values significantly less than 1).
- *Inference:* Upload speeds show a moderate to strong level of spatial autocorrelation.

- **Download Speed (`avg_d_speed_mbps`):**
  - Moran's I = 0.33 ⇒ Moderate clustering (weaker than upload speeds, but still significant).
  - Geary's C = 0.63 ⇒ Confirms clustering, though less pronounced.
  - *Inference:* Download speeds exhibit spatial patterns with more variability.
- **Latency (`avg_lat_ms`):**
  - Moran's I = 0.04 ⇒ Near-zero autocorrelation.
  - Geary's C = 0.93 ⇒ Close to 1, indicating randomness.
  - *Inference:* Latency values appear spatially random, with no clear geographic clustering.

2) Top-10 cities local statistics for download and upload: Why Global SAC (Spatial Autocorrelation) Measures Aren't Enough

Global measures like Moran's I and Geary's C assume **stationarity**, meaning spatial patterns are consistent across the entire study region. However, in real-world scenarios:

- Spatial patterns often vary by location (e.g., urban centers vs. rural areas).
- Localized trends can cancel each other out when aggregated globally.

Hence, we rely on **Local Indicators of Spatial Autocorrelation (LISA)** to uncover more nuanced patterns:

- **Local Moran's I:** Detects clusters (high-high, low-low) and spatial outliers (high-low, low-high).
- **Local Geary's C:** Measures local dissimilarity. Values < 1 indicate clustering, values > 1 suggest dispersion.
- **Local Getis-Ord G*:** Identifies *hotspots* (high G*) and *coldspots* (low G*).

Key Findings from the Figure 9

- **Local Moran's I (in decreasing order):** Cities like *Abū Şuwayr*, *Al Khahidah*, and *Cairo* show high positive values (1.0–2.0), indicating strong local clustering of high speeds.
  Interpretation: These are likely high-speed urban clusters with robust infrastructure.
- **Local Geary's C (in increasing order):** Cities such as *Shuhrâ al Khaşımah* and *Al Obişyah* report very low values (0.00–0.08), indicating tight homogeneity in download speeds.
  Interpretation: Consistency in performance across neighborhoods confirms stable local connectivity.
- **Local Getis-Ord G* (in decreasing order):** Cities like *Al 'Ashir min Ramaān* and *Cairo* emerge as hotspots.
  Interpretation: These regions significantly outperform surrounding areas in download speed.

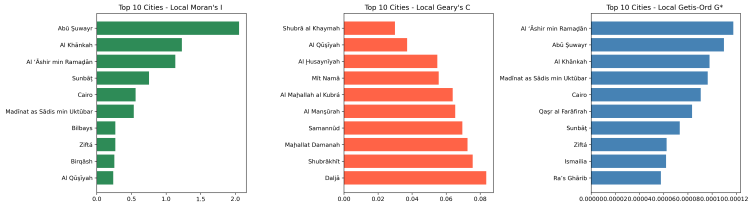

Fig. 9: Bar chart of local spatial statistics for the top 10 cities by download speed.

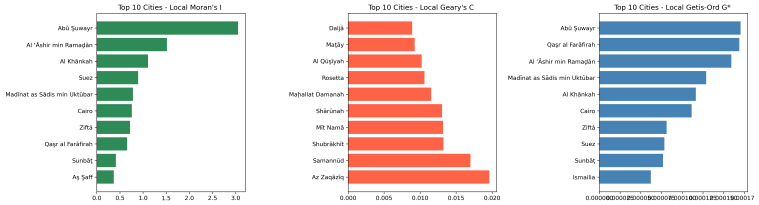

Fig. 10: Bar chart of local spatial statistics for the top 10 cities by upload speed.

Key Findings from the Figure 10

Upload speeds in Egypt exhibit even stronger spatial clustering compared to download speeds:

- Cities such as *Abū Şuwayr* (Moran's I = 3.05) and *Al 'Āshir min Ramaān* (I = 1.51) display extremely high Moran's I values, marking them as high-performance upload hotspots.
- Local Geary's C values are remarkably low (e.g., 0.0089), highlighting strong homogeneity and consistent performance within these clusters.
- Rural areas, in contrast, exhibit weaker clustering—pointing to an urban-rural digital divide.
- Getis-Ord G* statistics further confirm that urban centers benefit from centralized and superior infrastructure, forming statistically significant upload speed hotspots.

Note: All reported values are statistically significant ($p < 0.05$), ensuring that the observed spatial patterns are reliable and not due to random chance.

3) Local Moron Clustering: The Figure 11 and Figure 12 show clusters based on Local Moran's I values, which help us understand spatial patterns in the data. The first plot in Figure 11 shows different types of clusters: high-high, low-low, high-low, and low-high. While clusters are mostly visible around the central region of Egypt, the plot is cluttered due to the presence of many non-significant and noisy points. To make this clearer, the next plot in Figure 12 includes only statistically significant clusters (based on p-values). By removing the noisy and non-significant points, the real clusters stand out. In this cleaner plot, we can clearly see the high-high and low-low clusters, which indicate strong positive spatial autocorrelation, while the high-low and low-high clusters suggest negative spatial relationships.

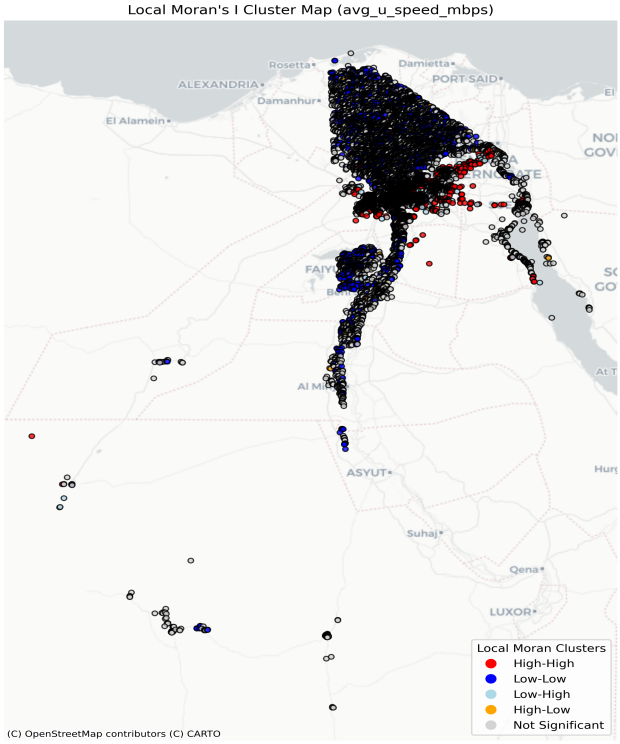

Fig. 11: Local Moran's I cluster map identifying statistically all spatial clusters and outliers.

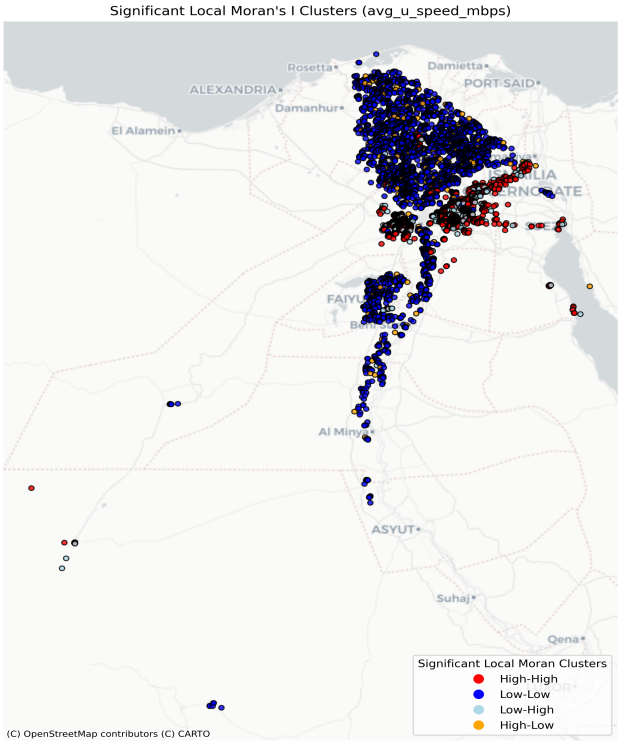

Fig. 12: Local Moran's I cluster map identifying statistically significant spatial clusters and outliers.

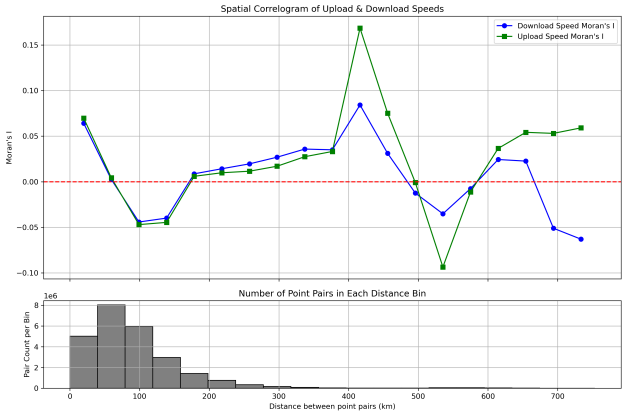

Fig. 13: Spatial correlogram showing autocorrelation strength at increasing distance lags.

4) Correlogram: The correlogram in the the Figure 13 shows how spatial autocorrelation, measured by Moran's I, changes with distance for both upload and download speeds. At shorter distances—especially up to around 50 km—both upload and download speeds display positive spatial autocorrelation. This means nearby locations tend to have similar network speeds. Interestingly, in the 75–125 km range, both variables show negative spatial autocorrelation, suggesting that locations at this distance are more likely to have contrasting speeds. A notable observation is that the highest positive autocorrelation for upload speeds occurs around 400 km (Moran's I 0.16), while for download speeds it peaks at the same distance with a smaller value (0.08). Beyond 500 km, download speeds show mostly negative spatial autocorrelation, whereas upload speeds fluctuate but generally remain positive. However, it's important to note that the majority of point pairs in the dataset are within 200 km. Because of this, the correlogram becomes more erratic at greater distances. Therefore, it is best to avoid drawing strong inferences beyond 200–250 km.

C. Spatial Heterogeneity

Spatial heterogeneity refers to the variation or differences in data values across different geographical locations. In simple terms, it means that things are not the same everywhere on the map.

It occurs because different areas may have unique characteristics—such as population density, infrastructure, environment, or economic activity, which influence the values of the data being measured

1) Chloropeth of local moron's I: As already seen in the previous sections, upload and download speeds shows partial spatial autocorrelation, suggesting that the data is not uniform across locations. To explore this further, we looked at local Geary's C values for different places and plotted them for the top 10 cities in increasing order. Our dataset includes 80 unique cities, and many of them show high local Geary's C values, which is a strong indicator of spatial heterogeneity—meaning nearby locations differ significantly in their network measurements. This finding is also supported

by local Moran's I values, where several locations have very low I values, indicating little similarity between neighboring points.

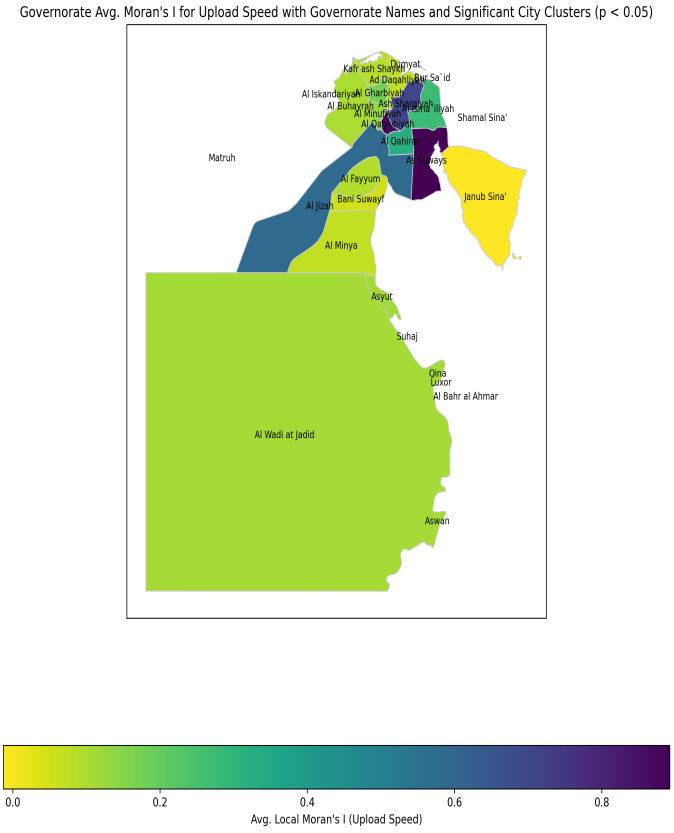

Fig. 14: Choropleth map of Local Moran's I values, highlighting areas with significant spatial clustering of similar download speed values.

To visualize this, we plotted a choropleth map in Figure 14 of Egypt based on average local Moran's I values for each governorate(Egypt has 27 governorates, similar to states, each containing multiple cities). In the map, the color gradient represents the strength of spatial autocorrelation—lighter colors (as seen in the color bar) indicate lower Moran's I values and therefore more heterogeneity, while darker regions like As-Suways and Al-Jizah show higher Moran's I values, suggesting better spatial autocorrelation.

Overall, it is clear that the data is spatially heterogeneous, which is common in many real-world geographic datasets.

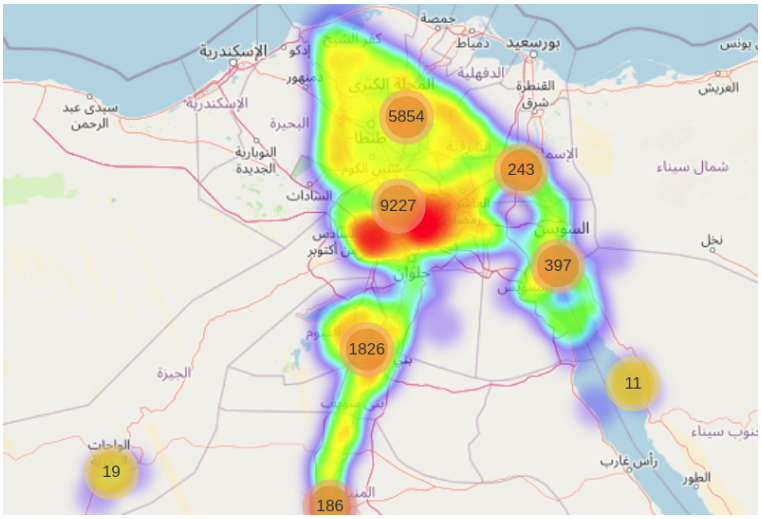

Fig. 15: Heatmap denoting average download speed across Egypt

2) Heat Map: The heatmap visualization in Figure 15 generated using Folium, combined with average download speed in megabits per second, provides a powerful spatial understanding of network performance across Egypt. Key observations are as follows:

Urban Hotspots of High Download Speed

- *Greater Cairo Region (9227 data points)* is the most prominent hotspot, showing intense red coloration, indicating both high density of measurements and high download speeds.
- Key contributors likely include *Giza, Nasr City*, and *6th of October*, due to their high population density and robust network infrastructure.
- *Tanta, El-Mahalla El-Kubra*, and other nearby Delta cities (5854 points) form another major hotspot, benefiting from relatively advanced infrastructure due to economic and population centrality.

Moderate Performance Zones

- *Suez (397 points)* and *Ismailia (243 points)* display a mix of green and yellow zones, reflecting moderate download speeds and medium measurement density.
- These cities represent a transition between high-performance urban centers and lower-speed rural or peripheral areas.

Upper Egypt Corridor

- A vertical cluster from *Cairo through Beni Suef, Minya, and Assiut* (1826 points) reveals a moderately high-performance zone.
- While not reaching Cairo's speeds, the green/yellow coloring indicates relatively consistent connectivity, likely due to the presence of major population centers along the Nile corridor.

Peripheral and Low-Coverage Areas

- The *Western Desert region* (e.g., Al Wahat, 19 points) and parts of *South Sinai* (11 points) show sparse data with low intensity.

- These areas appear in blue or have weak/no coloring, suggesting poor network coverage, fewer users, or very low download speeds.
- Their marginal status is supported by low measurement counts and geographic distance from infrastructure hubs.

Digital Divide Observation

- The heatmap clearly shows a *stark urban-rural divide*.
- *Urban centers like Cairo* dominate in both data density and download speed.
- *Rural and desert regions* lag behind in connectivity, underscoring significant infrastructure disparities.
- *Policy implication:* Strategic investments in underserved regions are essential to bridge the digital gap.

The heatmap provides a clear spatial understanding of network quality disparities across Egypt. While major cities enjoy robust download speeds and coverage, peripheral and rural areas lag behind.

D. Stationarity

Stationarity in spatial data means that the statistical properties of the data , like the mean, variance, and spatial relationships — stay the same across the entire area being studied. When data is non-stationary, patterns or values change from one location to another, which is common in real-world scenarios.

While, analyzing heterogeneity and spatial autocorrelation, we found that some areas have strong spatial autocorrelation while others show very weak or no autocorrelation. The choropleth maps also showed that the data is spatially heterogeneous — different regions have different behaviors. This clearly indicates that our data is not stationary, since the relationships between data points change across locations.

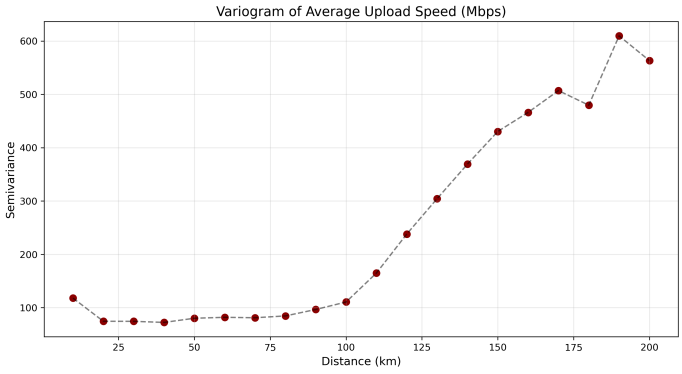

Fig. 16: Variogram displaying the semivariance of upload speeds with respect to spatial distance.

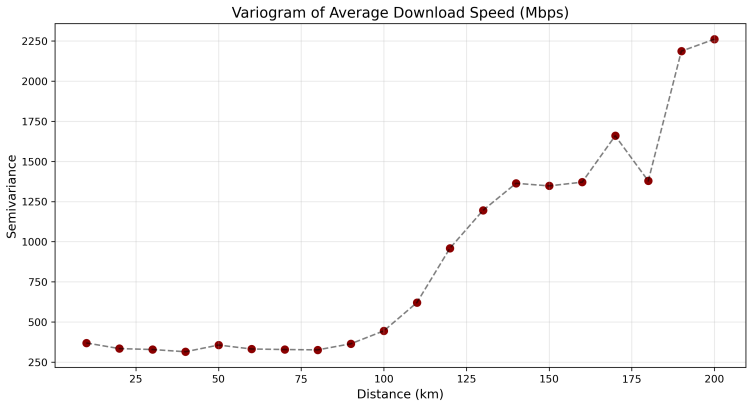

Fig. 17: Variogram displaying the semivariance of download speeds with respect to spatial distance.

To further support this, we plotted variograms for average upload in Figure 16 and download in Figure 17 speeds. In both variograms:

The semivariance is low and stable for short distances (up to about 100 km), meaning nearby locations tend to have similar values. After 100 km, the semivariance increases sharply, especially for download speeds, which continue to rise steeply up to 200 km. This pattern suggests that as distance increases, the difference between data points also increases, and the data does not stabilize over distance. This rising semivariance with no clear plateau further confirms that the data is spatially non-stationary.

IV. Spatial Data Mining

A. Deeper Analysis of the Dataset

In earlier sections, we looked at the distribution of variables and basic plots to understand what data we are working with and where the data points are located geographically. For instance, we analyzed the distribution of average download speeds across cities (see Figure 2).

Now, to uncover more meaningful insights, we zoom in on cities with the highest upload and download speeds. This helps us understand which areas perform best and if there are any spatial patterns behind high-performing regions.

As seen in the Figure 18 and 19:

- **For download speeds**, the top-performing cities are:
  - **Al-Ashir min Ramadan**
  - **Abu Suwayur**
  - **Al Khankah**
- **For upload speeds**, the cities with the highest averages are:
  - **Qasr al-Farafirah**
  - **Abu Suwayur**
  - **Al Khankah**

From the plots, we can observe that **Abu Suwayur** appears in both lists, suggesting it may have a consistently strong network infrastructure overall. These patterns give us early indicators of potential regional differences in infrastructure quality, user demand, or service provider distribution.

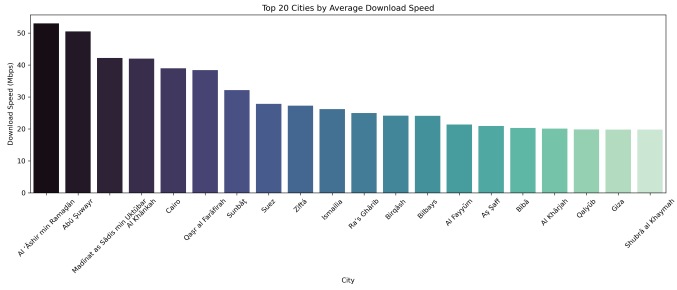

Fig. 18: Bar chart denoting Top 20 cities download speed in Egypt

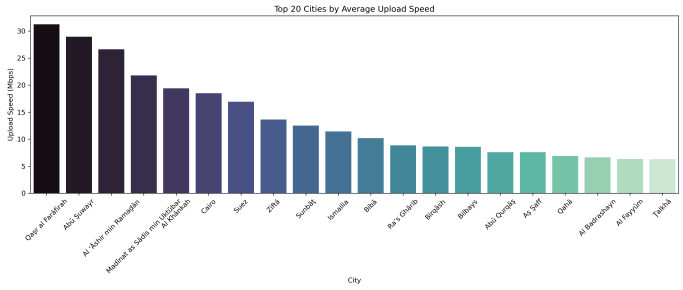

Fig. 19: Bar chart denoting Top 20 cities Upload speed in Egypt

1) Correlation Matrix: To check whether any of the numerical variables in our dataset are related to one another, we plotted a correlation matrix using five key numerical features: number of tests, number of devices, average download speed, average upload speed, and average latency.

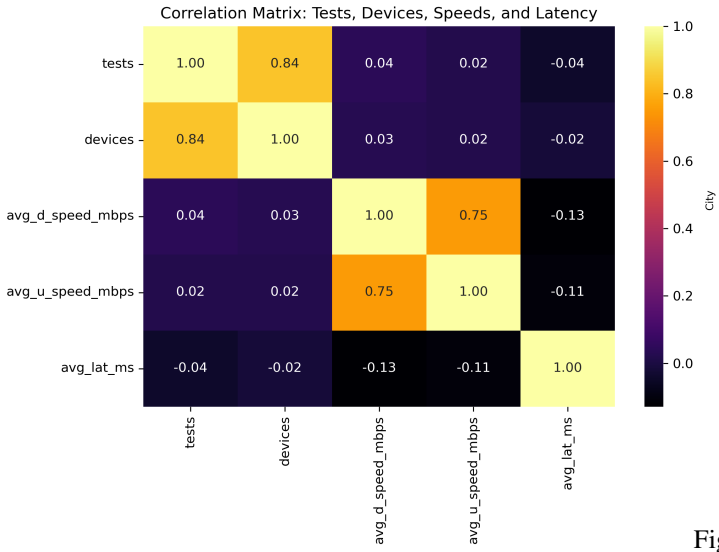

Fig. 20: Correlation matrix showing relationships between network performance metrics

Key Observations from the Correlation Matrix in Figure 20

- *Strong positive correlation between tests and devices (0.84):*
  Areas with more tests typically also have more devices. This is expected, as more devices lead to more speed tests being conducted.
- *High positive correlation between download and upload speeds (0.75):*
  Regions with faster download speeds also tend to have faster upload speeds. This reflects overall network quality trends, where both metrics improve or degrade together.
- *Very weak correlation between tests/devices and speeds:*
  The number of tests or devices in a region has almost no correlation with the download or upload speeds. This implies that simply having more activity (tests or devices) does not guarantee better performance.
- *Negative correlation between speeds and latency:*
  There is a weak negative correlation between download/upload speeds and latency (-0.13 and -0.11 respectively). As expected, higher speeds generally coincide with lower latency, but the relationship is not very strong.
- *No significant correlation between latency and tests/devices:*
  Latency does not appear to be meaningfully linked to how many devices or tests exist in a region, indicating that other factors influence latency more significantly.

These findings help us understand the internal relationships in the data and provide a foundation for more advanced spatial analysis and clustering in the upcoming sections.

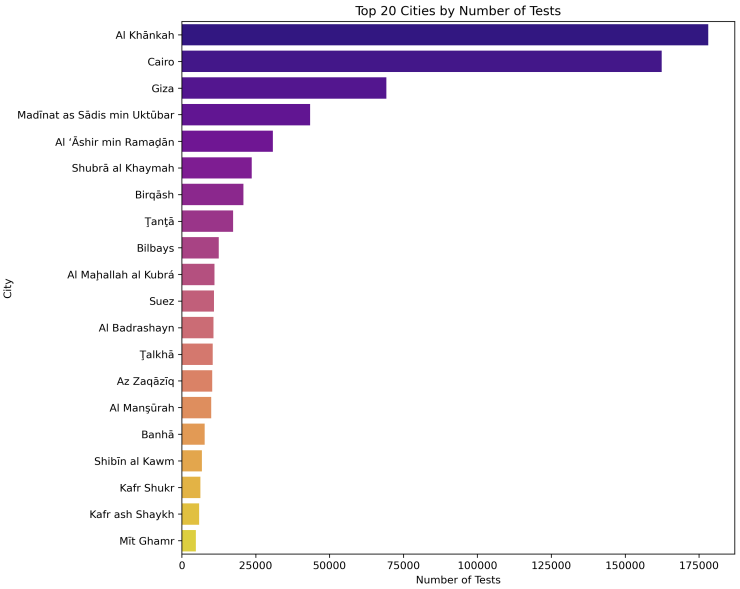

Fig. 21: Bar chart denoting Top 20 cities by number of tests in Egypt

2) Bar Chart-Top 20 Cities by Number of Tests: This bar chart in Figure 21 shows the top 20 cities in Egypt based on the number of network tests conducted.

The cities Al Khānkah, Cairo, and Giza clearly lead the chart, with a significantly higher number of tests compared to others. These three cities alone account for a large portion of all the tests, likely due to their large populations and higher demand for internet performance.

Other cities like Madinat as Sādis min Uktūbar, Al ʻĀshir min Ramaḍān, and Shubrā al Khaymah also show noticeable testing activity but fall behind the top three. The number of tests gradually drops as we go down the list.

Inference: Most network tests are performed in Egypt's largest and most urbanized cities, suggesting that densely populated areas are more engaged in network usage and monitoring. Smaller towns and rural regions are less represented in testing data.

3) Feature Engineering- Adding a Governorate Column: Egypt is divided into 27 governorates, each consisting of several cities. In our dataset, we had 80 unique cities, but no direct information about which governorate each city belongs to. Knowing the governorate for each test record would help us create better grouped (aggregated) visualizations, especially maps and bar charts.

To solve this, we created a new feature: **Governorate**.

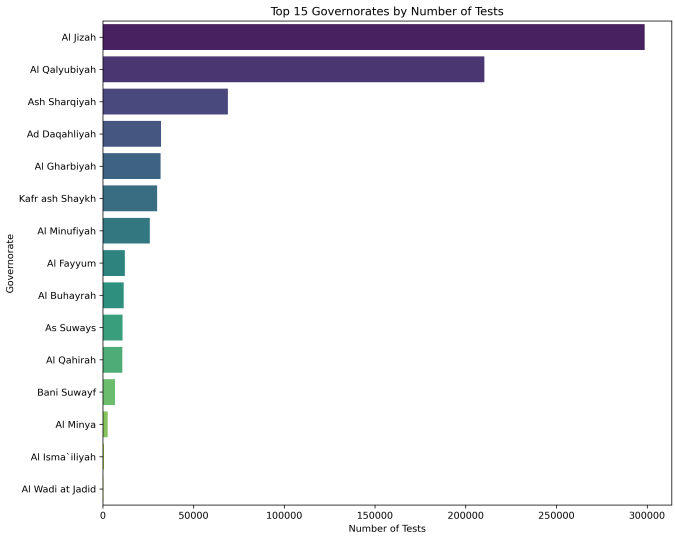

Fig. 22: Bar chart dentoting Top 15 Governorates by number of tests in Egypt

Here's how it was done:

- We used a shapefile of Egypt's administrative boundaries [2], which included the geographic coordinates of all 27 governorates.
- For each city in our dataset, we checked whether its latitude and longitude fell within the boundary of any governorate.
- Once a match was found, we assigned the appropriate governorate to that city and added it as a new column in our dataset.
- This feature proved extremely useful. For example, it enabled us to plot **choropleth maps** showing how tests are distributed across Egypt's governorates.

4) Bar Chart- Top 15 Governorates by Number of Tests: This bar chart Figure 22 shows the top 15 governorates in Egypt based on the number of network tests.

The governorates Giza (Al Jizah), Qalyubiyah (Al Qalyubiyah), and Sharqia (Ash Sharqiyah) have the highest number of tests. Giza and Qalyubiyah, both part of the Greater Cairo metropolitan region, lead by a wide margin. These areas are highly urbanized, economically important, and densely populated, which explains the high testing activity.

At the lower end of the top 15 are governorates like Al Wadi al Jadid, Ismailia (Al Isma'iliyah), and Minya (Al Minya)with fewer tests. These are more rural or less populated areas with lower internet penetration or less interest in testing.

Inference:
Testing activity follows population and development patterns—urban and economically strong regions perform more tests, while rural areas show lower engagement.

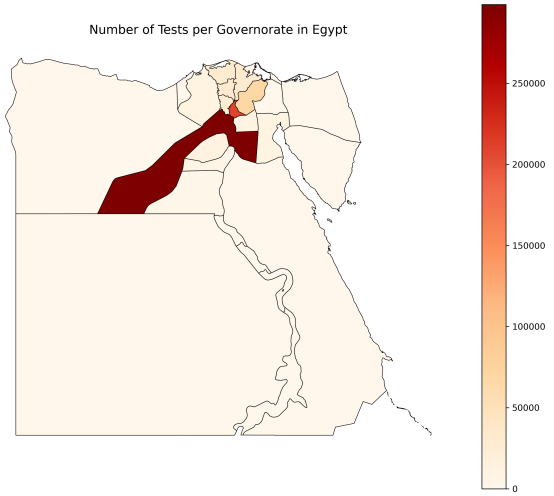

Fig. 23: Chloropleth map of number of tests per Governorate in Egypt

5) Choropleth Map- Tests per Governorate: We used the new Governorate feature to create a choropleth map that shows where the tests are happening across Egypt.

In this map shown in Figure 23:

- Darker colors represent governorates with more tests.
- Lighter colors show areas with fewer tests.
- From the map, it is clear that northern Egypt, especially the Greater Cairo region (Cairo, Giza, Qalyubiyah), has the highest concentration of tests. These areas are urban, densely populated, and have better network infrastructure.

Inference: There is a clear regional difference in testing activity—urban north dominates in terms of both population

and internet testing, while southern and rural governorates have significantly fewer tests.

B. Outlier Detection

An outlier is a data point that significantly differs from other observations in a dataset. It lies far outside the typical range of the data and may indicate variability in measurement, experimental error, or a novel insight.

In our dataset, we are dealing with two types of outliers:

- Global Outliers
- Spatial Outliers

1) Global Outliers: These are data points that are significantly different from the rest of the data when looking at the entire dataset, regardless of where they are located geographically. For example, a download speed of 150 Mbps in a dataset where most values are below 20 Mbps would be considered a global outlier.

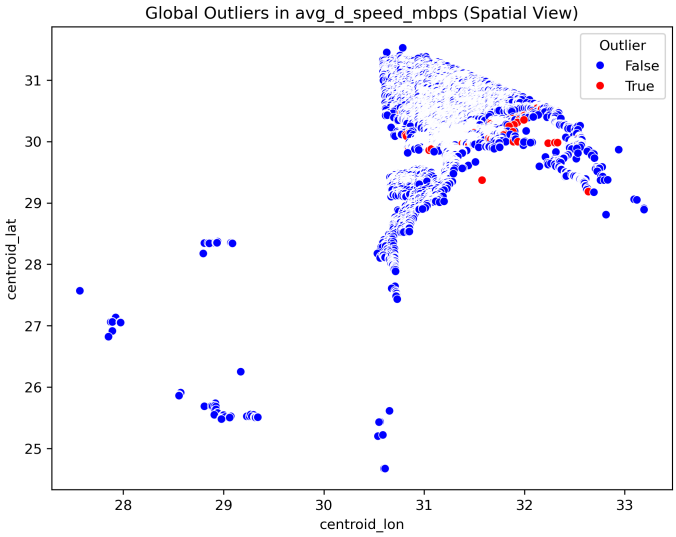

Fig. 24: Spatial distribution of global outliers in average download speed (avg_d_speed_mbps).

In this section, we are identifying global outliers in the average download speed across Egypt. The same approach can be applied to other performance metrics such as upload speed or latency, especially since upload and download speeds were found to be highly correlated in our earlier analysis.

To detect global outliers, we used the Z-score method. This method standardizes the data to determine how many standard deviations each value is from the mean. Points with a Z-score greater than ±3 are typically considered outliers.

We first visualized the distribution using a boxplot in Figure 4 to check for any data points that lie outside the typical range. However, since most of the data in Egypt has relatively low download speeds, the few data points with very high speeds appeared as outliers. These are likely to be locations with better infrastructure or newer technology deployment.

To better understand where these outliers are geographically located, we plotted them on a scatterplot based on their spatial coordinates (latitude and longitude). In the plot below, the red dots represent outliers—places with significantly higher download speeds than the rest of the country—while blue dots represent normal values.

This spatial view helps us identify the geographical clusters or isolated high-speed areas, providing insights into where network performance is exceptionally good.

2) Spatial Outliers: These are data points that are unusual relative to their geographic neighbors. A value might not seem extreme when viewed across the entire dataset, but it could stand out when compared only to nearby points. For example, a city with very low latency surrounded by high-latency cities could be a spatial outlier.

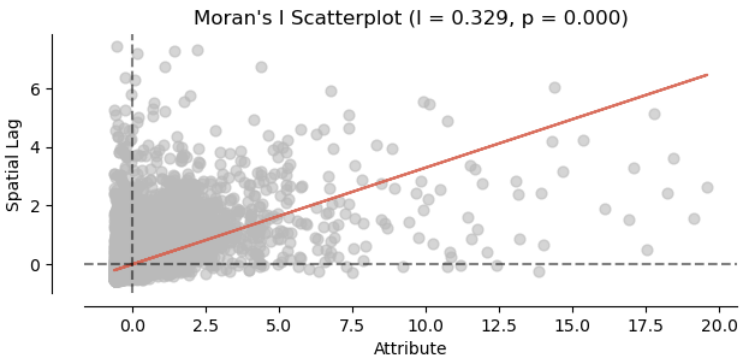

Fig. 25: Moran's I scatterplot showing spatial autocorrelation of network performance (I = 0.329, p = 0.000).

To identify spatial outliers, we used Moran's I statistic, a measure of spatial autocorrelation that indicates whether similar values cluster together geographically.

The Moran's I value of 0.329 with a p-value of 0.000 indicates a statistically significant positive spatial autocorrelation, meaning that cities with similar average download speeds tend to be located near each other.

Scatterplot Explanation:

- **X-axis:** Actual values (e.g., average download speed).
- **Y-axis:** Spatial lag (the average value of neighboring locations).
- **Red regression line:** Indicates the global trend of spatial association.

Interpretation:

- Points in the **upper-right** and **lower-left** quadrants show high-high and low-low clusters, i.e., spatial clusters with similar values.
- Points in the **upper-left** or **lower-right** quadrants indicate **spatial outliers**, where a value is very different from its neighbors (e.g., a city with high speed surrounded by low-speed areas).

These spatial outliers are crucial for identifying regions with unexpectedly high or low performance relative to their neighbors.

C. Spatial Clustering

Clustering is a process of grouping a set of objects such that intra-group similarity is maximized, and inter-group similarity is minimized. In the context of spatial data, this means identifying regions or areas where certain patterns—such as network performance—are similar within but different from others.

Before applying clustering algorithms, it is essential to check whether the data exhibits any spatial pattern or if it is just randomly distributed. This is where the concept of Complete Spatial Randomness (CSR) becomes important. Complete Spatial Randomness assumes that events (or data points) are uniformly and independently distributed across space.

1) Ripley's K Test to Check for Spatial Clustering: To examine this, we used Ripley's K function, a spatial statistical test used to detect departures from CSR. It helps assess whether a point pattern is random, clustered, or dispersed over a range of distances.

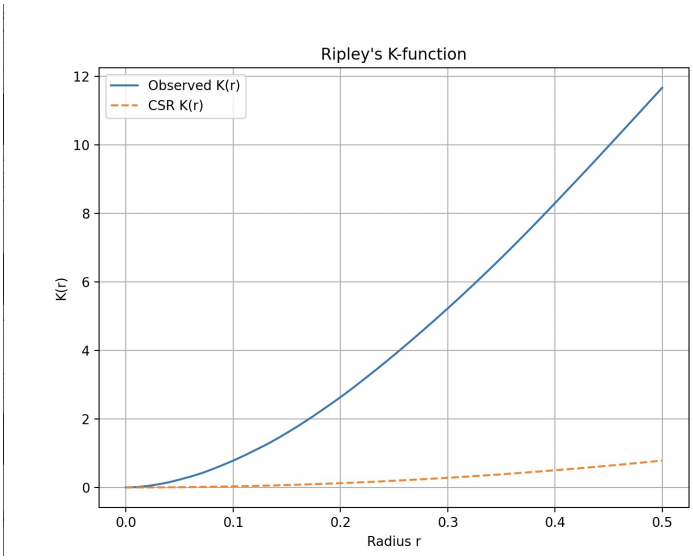

Fig. 26: Ripley's K-function analysis of spatial point patterns in network performance measurements. The plot compares the observed K(r) (solid line) against complete spatial randomness (CSR, dashed line) across radii (0.0-0.5 units).

How to Read the Ripley's K Plot:

- Observed K(r) – Blue Line: Represents the actual K-values from the dataset at different spatial scales (r).
- CSR K(r) – Orange Dashed Line: Represents expected K-values under the null hypothesis of CSR.

Interpretation:
Clustering: The blue line lies significantly above the orange line at all distances. This means that at every spatial scale, there are more data points clustered together than would be expected by random chance.
The Ripley's K analysis strongly indicates that our point data is spatially clustered across all scales. This validates the use of spatial clustering algorithms to identify meaningful groupings in the dataset.

2) Clustering Using DBSCAN: After confirming that our spatial data exhibits strong clustering behavior (as shown through Ripley's K test), we proceeded with applying clustering algorithms to identify distinct regions of similar network performance. For this task, we chose the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm.

DBSCAN was a logical and effective choice for several reasons. Unlike K-Means or K-Medoids, DBSCAN does not require us to predefine the number of clusters, which is particularly useful here since we do not have prior knowledge of how many meaningful groupings exist in the data. Additionally, DBSCAN is well-suited for spatial data as it groups together points that are densely packed, allowing for clusters of arbitrary shape and size. Another important advantage is its ability to identify and label outliers or noise points, which is critical in real-world data where anomalies or sparsely populated areas are common. These features make DBSCAN an ideal choice for our use case.

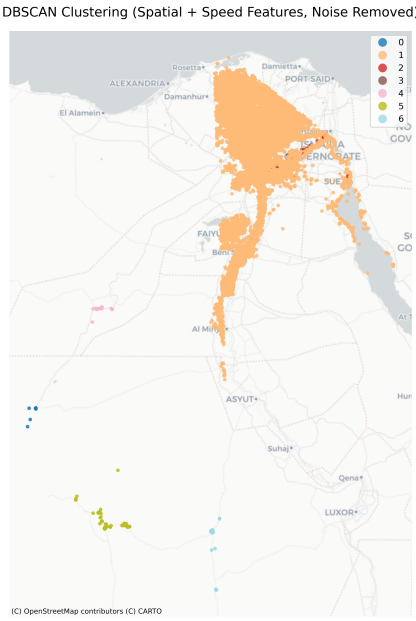

Fig. 27: DBSCAN clustering results for network performance data using spatial and speed features

Initially, we tried clustering based only on spatial coordinates, specifically latitude and longitude. While this created clusters based on geographical proximity, it ignored crucial performance-related differences between nearby areas. For example, two adjacent regions could have drastically different download or upload speeds, and clustering them together would be misleading. Realizing this, we enhanced our clustering input by incorporating additional features such as average download speed, average upload speed, and the number of speed tests conducted in each area. This ensured that clusters

reflected not only spatial closeness but also similarity in network performance.

Using DBSCAN with these combined features, we obtained meaningful groupings that captured both spatial and technical attributes. The algorithm identified seven distinct clusters, excluding 57 noise points that did not belong to any cluster based on the density threshold. These clusters showed significant variation in performance characteristics. For instance, some clusters had extremely high average speeds and very few tests, indicating areas with potentially high-capacity but low-usage infrastructure. Others had moderate speeds but a large number of tests, suggesting more populated or actively used regions.

Inference:

Cluster 0: This group exhibited moderate performance:

- Average download speed: **17.61 Mbps**
- Average upload speed: **7.16 Mbps**
- Tests per location: **2.38**

This likely represents regions with stable but mid-level connectivity.

Cluster 1: Areas in this cluster showed slightly higher performance:

- Average download speed: **23.18 Mbps**
- Average upload speed: **8.61 Mbps**
- Tests per location: **41.0**

These are likely densely populated or actively used areas with relatively good infrastructure.

Cluster 2: This cluster stood out due to its very high performance:

- Average download speed: **253.70 Mbps**
- Average upload speed: **274.55 Mbps**
- Tests per location: **2**

This might indicate areas with premium internet services or specialized usage zones.

Cluster 3: Another high-performance cluster:

- Average download speed: **728.25 Mbps**
- Average upload speed: **98.38 Mbps**
- Tests per location: **1.0**

This could represent a highly specialized or limited-use region where extremely fast internet is available but not widely tested.

Cluster 4: This group displayed low network performance:

- Average download speed: **13.20 Mbps**
- Average upload speed: **2.46 Mbps**
- Tests per location: **2.89**

Cluster 5: Similar to Cluster 0, this cluster had moderate speeds:

- Average download speed: **16.25 Mbps**
- Average upload speed: **4.67 Mbps**
- Tests per location: **4.14**

Cluster 6: This final group showed good performance:

- Average download speed: **20.10 Mbps**
- Average upload speed: **5.34 Mbps**
- Tests per location: **5.15**

These areas might have reliable infrastructure and moderately engaged users.

D. Spatial Colocation

Spatial colocation analysis is a way to see if different types of things tend to be found in the same places. It's not just about whether two things are related, but whether their locations are related. This helps us understand how things are connected in space.

In this study, we looked at whether high-speed internet access is found in certain types of areas. We wanted to see if access is spread evenly, or if it's concentrated in some areas more than others. We thought that internet speed might be related to things like how many people live in an area, or the presence of important community resources.

1) Experiment 1: High Population Density Areas: First, we looked at the relationship between fast internet and areas with lots of people. We used a map of populated places in Egypt to represent population density. We called internet access "fast" if it was faster than the middle speed in our data. Then, we checked if these fast internet locations were close to populated places (within 1.5 km).

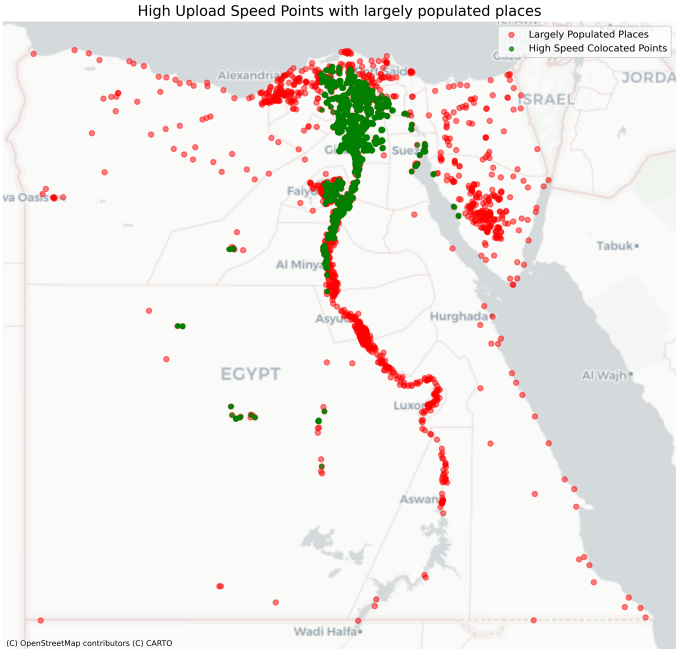

Fig. 28: Geographic colocation of high upload speed points with populated areas in Egypt

The map in Figure 28 showed that fast internet access is colocated with high population density areas in several key regions, suggesting a degree of success in infrastructure deployment.

- Greater Cairo: Has the most fast internet, showing it has the best internet infrastructure and a strong alignment with population centers.
- Nile Delta: Shows good fast internet coverage, indicating a positive relationship with population density, though less concentrated than in Cairo.

However, some disparities remain:

- Upper Egypt (South): Shows less fast internet the farther you get from Cairo, even though there are still many populated places.
- Coastal and Desert Regions: Have many populated places but very little fast internet.

2) Experiment 2: Proximity to Educational Institutes: We also examined the relationship between high-speed internet access and proximity to educational institutions. We used a shapefile of Egypt's educational centers, obtained from [3], and again employed the same colocation methodology, checking for fast internet locations within 1.5 km of these centers. The results of this experiment showed a weaker colocation pattern. This weaker relationship may be attributed to a significant number of missing values in the educational center shapefile, which limited the data available for analysis.

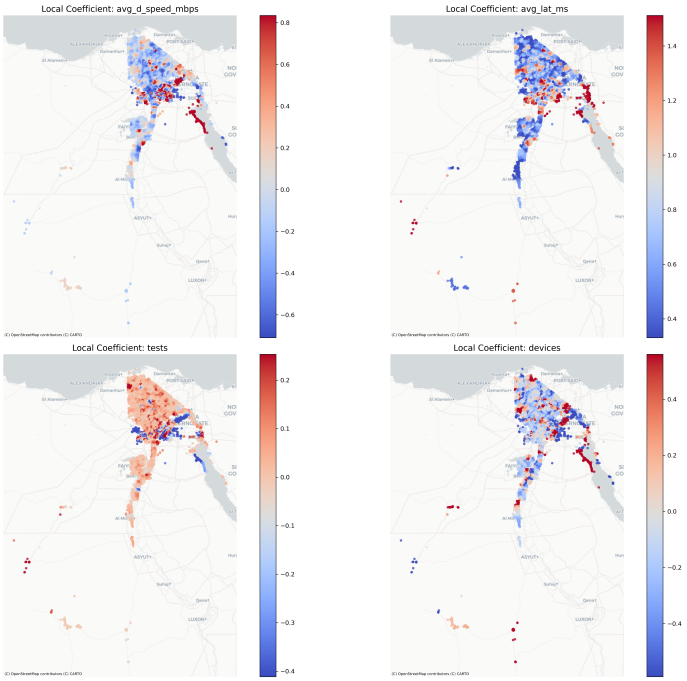

Fig. 29: GWR coefficients

E. Geographic Weighted Regression (GWR)

Geographically Weighted Regression (GWR) is a spatial analysis technique that extends traditional regression by allowing the relationships between dependent and independent variables to vary across geographic space. Unlike standard models that assume a global relationship, GWR captures local variations, making it ideal for understanding how predictors influence the outcome differently in different regions.

In this analysis, we applied GWR to model and understand the variation or predict average upload speed (Mbps) based on spatial and performance-related predictors. To achieve this, we first cleaned and geocoded our dataset, projecting the coordinates to a metric system suitable for spatial modeling. Our predictor variables included average download speed, latency, number of tests, and number of devices—the only useful features available in our limited dataset. These were standardized along with the target variable (upload speed), and weighted by the square root of the number of tests to give more influence to reliable measurements.

Using the mgwr Python package, we selected an optimal bandwidth using cross-validation and then fitted the GWR model. The resulting local coefficients for each predictor were mapped over Egypt using a basemap, visually illustrating how each factor's influence on upload speed varied across the country. Residual analysis was also conducted to assess model fit, revealing spatial patterns in prediction errors.

Despite having a small number of features, this approach allowed us to uncover important spatial heterogeneity in upload speed performance—insights that a global model would likely have missed.

Inference

1) **Model Performance: GWR vs. Global Regression**
The Geographically Weighted Regression (GWR) model significantly outperforms the traditional global regression (OLS) model in explaining the variation in upload speeds across Egypt. While the global model achieves an R^2 of 0.788 and an AIC of 78,368, the GWR model improves the explanatory power with an R^2 of 0.920 and a much lower AIC of 65,954. Furthermore, the residual sum of squares drops drastically from 83,918 in the global model to 31,643 in the GWR model, indicating a more accurate fit. The adaptive bandwidth selected for GWR was 89, meaning each local regression used the 89 nearest neighbors. These results clearly suggest that accounting for spatial heterogeneity substantially enhances the model's performance.
2) **Interpretation of Coefficient Maps**
The GWR coefficient maps reveal how the influence of each predictor on upload speed varies geographically:
label=.
   a) *Download Speed (avg_d_speed_mbps):*
   Most of the Nile Delta and Cairo regions exhibit positive coefficients, indicating that in these areas, higher download speeds are associated with higher upload speeds. In contrast, regions south of Cairo show weaker or negative relationships, suggesting local network dynamics differ significantly.
   b) *Latency (avg_lat_ms):*
   As expected, northern regions with higher latency show negative coefficients—higher latency correlates with lower upload speeds. However, in some southern and peripheral areas, coefficients

are positive or near zero, which could reflect local anomalies or noise in measurement.

c) *Number of Tests (tests):*
Positive coefficients in urban and central regions suggest that areas with more test activity tend to have better upload performance, possibly due to better infrastructure or more engaged users. Conversely, some remote regions show a negative trend, which might indicate frequent testing in areas of poor service.

d) *Number of Devices (devices):*
A mixed pattern emerges: in wealthier or more urbanized areas, more devices correlate with higher upload speeds (positive coefficients), while in certain congested or underdeveloped regions, more devices seem to result in lower performance (negative coefficients), possibly due to bandwidth sharing or network strain.

3) **Summary Statistics of GWR Coefficients**
The coefficients across all predictors exhibit substantial spatial variation. For example, the average coefficient for download speed is 0.83, with a standard deviation of 0.38, ranging from as low as 0.006 to as high as 2.62. Such wide variation, particularly for features like the number of devices, highlights the importance of local context in network performance modeling.

4) **Overall Conclusion**
The GWR model provides a far superior fit compared to the global regression model, clearly demonstrating that spatial variability plays a crucial role in explaining upload speed patterns across Egypt. These insights underline the value of localized modeling in understanding and improving telecommunication infrastructure, allowing for more targeted and region-specific interventions.

V. Conclusion

Through this comprehensive spatial analysis of internet speed and performance in Egypt, we began by exploring the distribution of key variables—particularly average upload and download speeds—to establish a foundational understanding of the dataset. Our investigation into spatial statistics, including global and local spatial autocorrelation (Moran's I, LISA, and Geary's C), revealed varying patterns of clustering, with some regions exhibiting high spatial dependence while others showed weaker associations. The presence of spatial heterogeneity was evident, aligning with the inherent variability in internet infrastructure and service quality across different locations. Further analysis through correlograms, choropleths, and variograms confirmed non-stationarity in the data, reinforcing the need for localized modeling approaches.

To deepen our insights, we engineered new features, identified global and spatial outliers, and employed spatial clustering technique (DBSCAN) to group regions with similar performance characteristics. Spatial colocation analysis uncovered meaningful patterns, while geographically weighted regression (GWR) proved highly effective in modeling upload speed variations, outperforming global models due to its ability to account for regional disparities. The success of GWR underscored the importance of localized analysis for heterogeneous datasets like ours.

In summary, this study not only highlighted the spatial dynamics of internet performance in Egypt but also demonstrated the value of advanced geospatial techniques in uncovering nuanced patterns and improving predictive accuracy. Future work could expand on these findings by incorporating additional covariates or temporal trends to further refine our understanding of connectivity landscapes.

VI. Implementation

A. Python Libraries Used

The entire code for visualizations, spatial statistics, and modeling was implemented in Python using the following key libraries:

- **NumPy** – Numerical computations and array operations.
- **Pandas** – Data manipulation, cleaning, and preprocessing.
- **Seaborn** & **Matplotlib** – Statistical visualizations and plotting.
- **Folium** – Interactive map visualizations for spatial point distributions.
- **GeoPandas** – Handling geospatial data and spatial operations.
- **Contextily** – Adding basemaps (e.g., OpenStreetMap) to geospatial plots.
- **MGWR** – Geographically Weighted Regression (GWR) for spatial modeling.
- **Libpysal** & **ESDA** – Computing spatial autocorrelation (Moran's I, LISA, Geary's C).
- **Shapely** – Geometric operations and spatial relationships.
- **SciPy** – Statistical tests and distance computations.
- **Scikit-learn** – DBSCAN clustering for spatial pattern detection.
- **OSMnx** – Accessing and processing OpenStreetMap data.

B. System Configuration

All code was executed in a Google Colab Notebook, leveraging cloud-based computational resources to efficiently handle large datasets and memory-intensive spatial operations.

C. Challenges Faced & Solutions

1) **Limited Relevant Columns in Dataset**
Problem: Only two columns (upload and download speed) were directly relevant.
Solution: Feature engineering introduced additional variables like governate-level aggregations and population density.

2) **Difficulty in Finding Egypt Shapefiles**
Problem: Many administrative boundary shapefiles were corrupted or incomplete.
Solution: A reliable shapefile was sourced from SimpleMaps.com for mapping accuracy.

3) **Lack of Complementary Datasets**
   Problem: Supplementary datasets (e.g., infrastructure, ISP coverage) had missing or sparse data.
   Solution: Used population data as a proxy, though ideal colocation analysis remained constrained.
4) **Computational Limitations for Variograms/Correlograms**
   Problem: Spatial autocorrelation tests on the full dataset (17,000+ points) caused crashes.
   Solution: Performed analysis on a representative subset to balance performance and accuracy.
5) **QGIS Installation Issues**
   Problem: Errors encountered while attempting to create custom shapefiles in QGIS.
   Solution: Relied on existing shapefiles and Python-based geospatial processing using GeoPandas.

VII. Contributions

This research represents a collaborative effort between both authors, with each contributing substantially to the exploratory data analysis, spatial statistical methods, and spatial data mining components. Our complementary expertise enabled a comprehensive investigation of internet speed patterns across Egypt.

Joint Contributions

- Exploratory Data Analysis (EDA): Collaboratively performed initial data exploration including distribution analysis, outlier detection, and feature engineering to prepare the dataset for spatial analysis.
- Methodological Development: Jointly designed the analytical framework and workflow for the spatial analysis pipeline.
- Report Composition: Co-authored all sections of the final report, ensuring cohesive integration of findings and consistent narrative flow.

Individual Contributions

Ullas G (IMT2022125)

- Designed and implemented Local Indicators of Spatial Association (LISA) mapping to identify significant local clusters of internet performance.
- Conducted spatial heterogeneity analysis through choropleth visualization and local statistics.
- Performed variogram analysis to examine spatial dependence structures and stationarity.
- Developed and optimized Geographically Weighted Regression (GWR) models to capture location-specific relationships.
- Executed spatial clustering algorithms (DBSCAN) to identify regions with similar connectivity profiles.

Mallikarjun (IMT2022116)

- Implemented global spatial autocorrelation measures (Moran's I, Geary's C) to assess overall spatial patterns.
- Conducted hotspot/coldspot analysis using Getis-Ord Gi* statistics.
- Generated and interpreted spatial correlograms to quantify scale-dependent autocorrelation.
- Performed outlier detection at both global and local scales to ensure data integrity.
- Implemented spatial colocation analysis to detect relationships between internet speeds and supplementary datasets.

This synergistic division of responsibilities allowed for a rigorous, multi-scale investigation of Egypt's internet infrastructure. By combining global pattern detection with localized modeling approaches, we ensured both comprehensive coverage and detailed spatial insights. Both authors contributed equally to troubleshooting computational challenges, refining methodologies, and synthesizing the final interpretations.

References

1) Mohamed Tag Eldin, Egypt Network Performance Data (Q4 2023), Kaggle. Available at: https://www.kaggle.com/datasets/mohamedtag04/egypt-network-performance-data-q4-2023
2) SimpleMaps, Egypt Administrative Regions Shapefile. Available at: https://simplemaps.com/gis/country/eg
3) Humanitarian Data Exchange (HDX), Egypt High Population Places (HOTOSM). Available at: https://data.humdata.org/dataset/hotosm$_egy_populated_places$