Exercise 2

Yumeng Lu

Problem 1

Task 1.1

1. **Images of the K-function charts for at least 2 different values for the Number of distance bands-parameters (1p)**

   Answer:

   - Images of the K-function charts for the Number of distance bands 50:

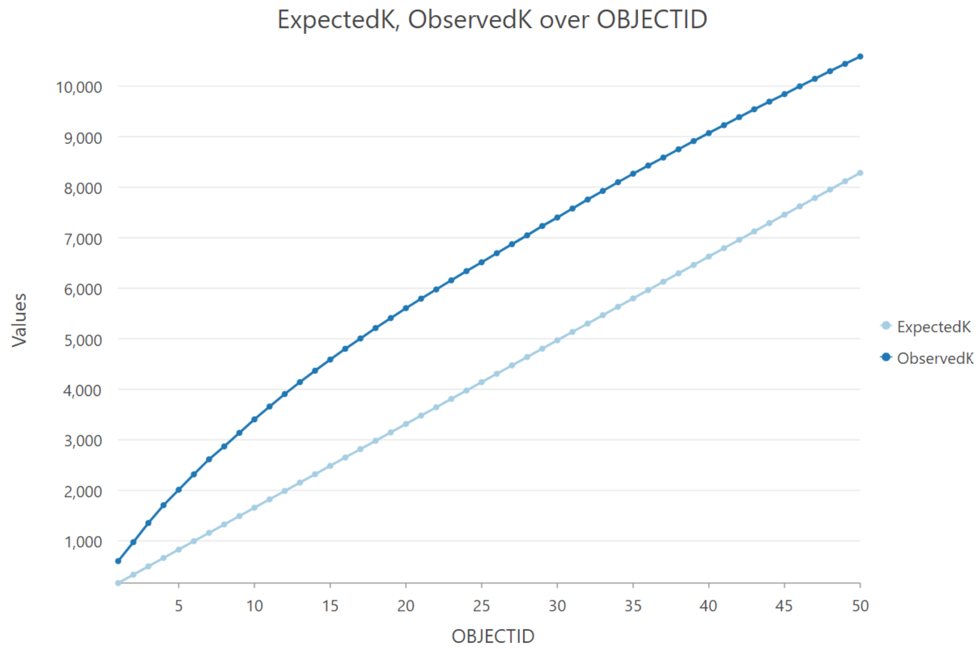

Fig 1 ExpectedK, ObservedK over OBJECTID (i.e., Number of distance bands 50)

   - Images of the K-function charts for the Number of distance bands 10:

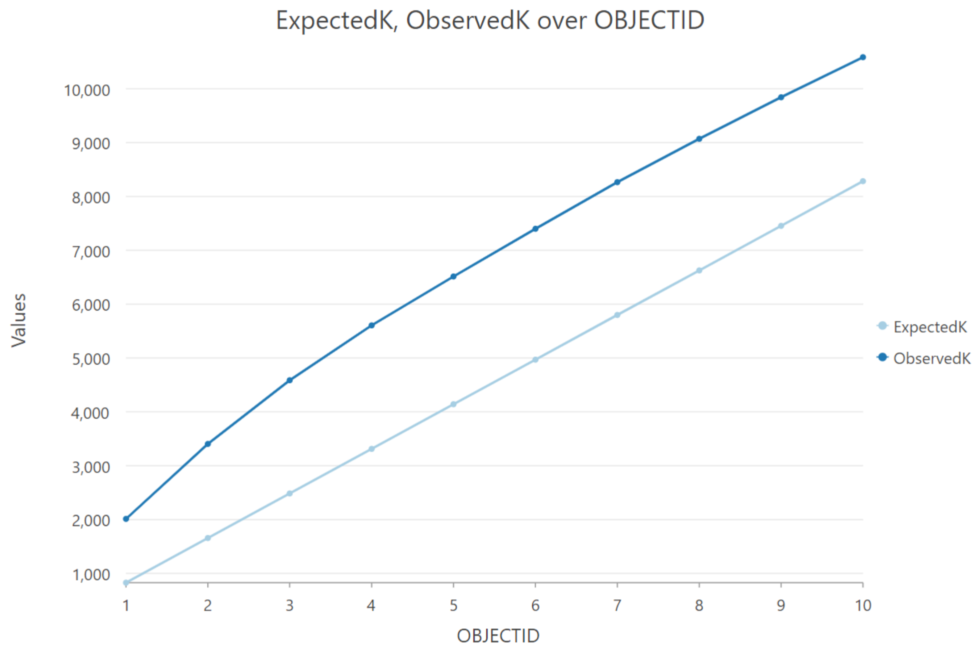

Fig 2 ExpectedK, ObservedK over OBJECTID (i.e., Number of distance bands 10)

2. Images of Kernel Density results with at least 2 different output cell sizes with the same search radius (1p)

Answer:

- Images of Kernel Density results with the same 'Search Radius 2400':

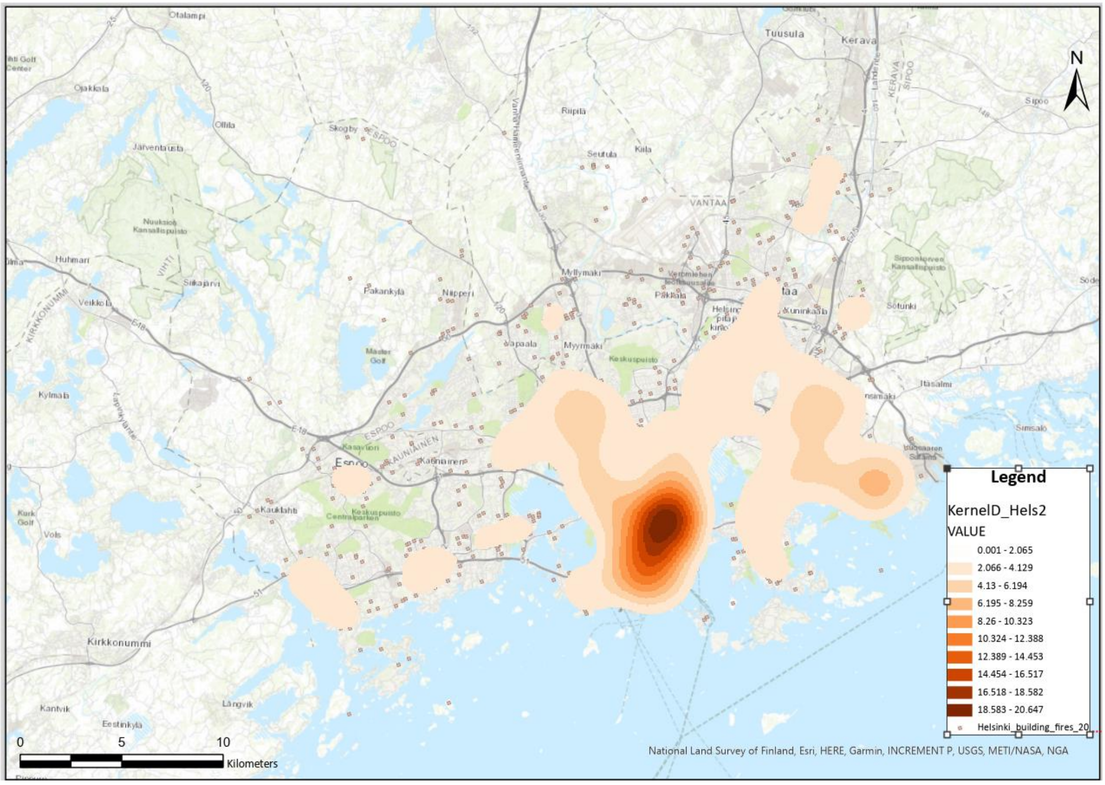

Fig 5 Kernel Density results with 'Output cell size 70' and 'Search Radius 2400'

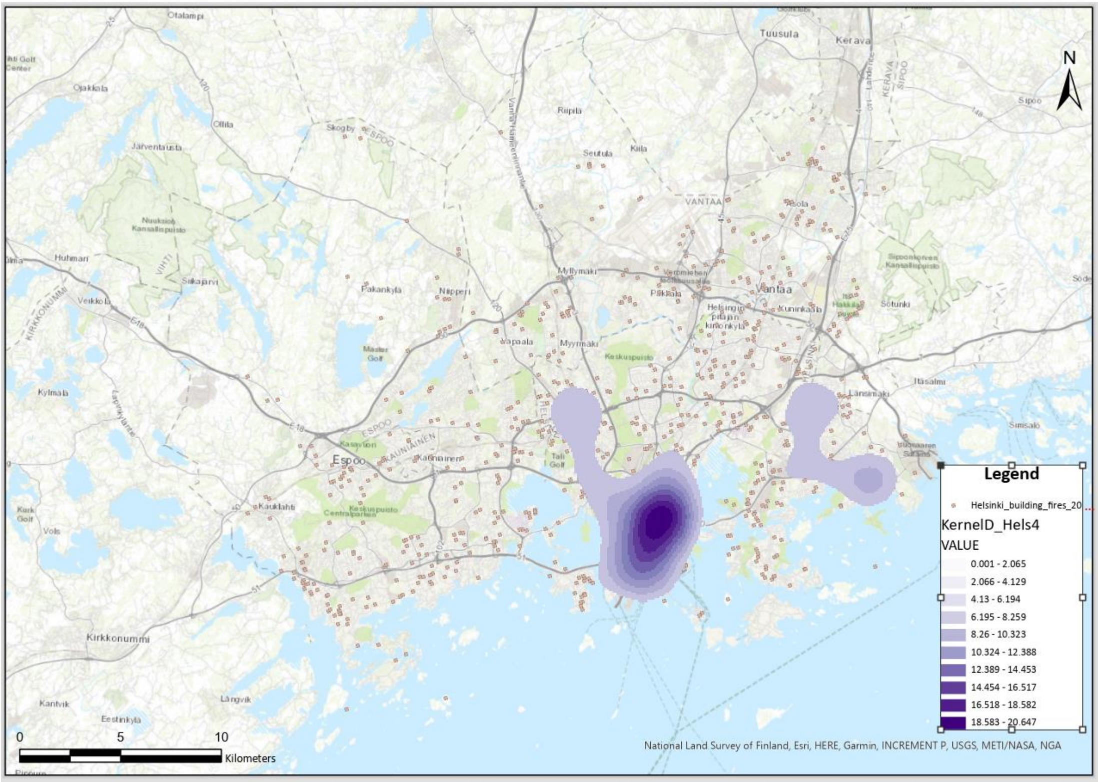

Fig 6 Kernel Density results with 'Output cell size 90' and 'Search Radius 2400'

3. Images of Kernel Density results with at least 3 different search radii with the same cell size (1p)

Answer:

- Images of Kernel Density results with the same ‘Output cell size 70’:

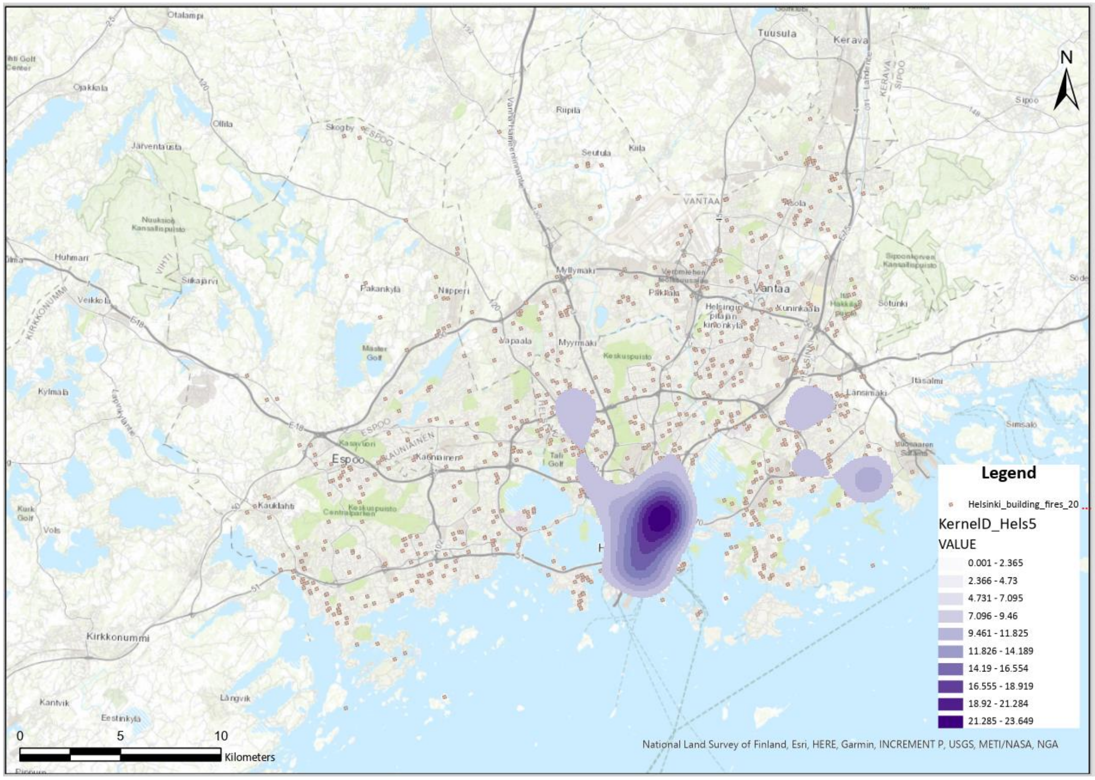

Fig 7 Kernel Density results with ‘Output cell size 70’ and ‘Search Radius 2100’

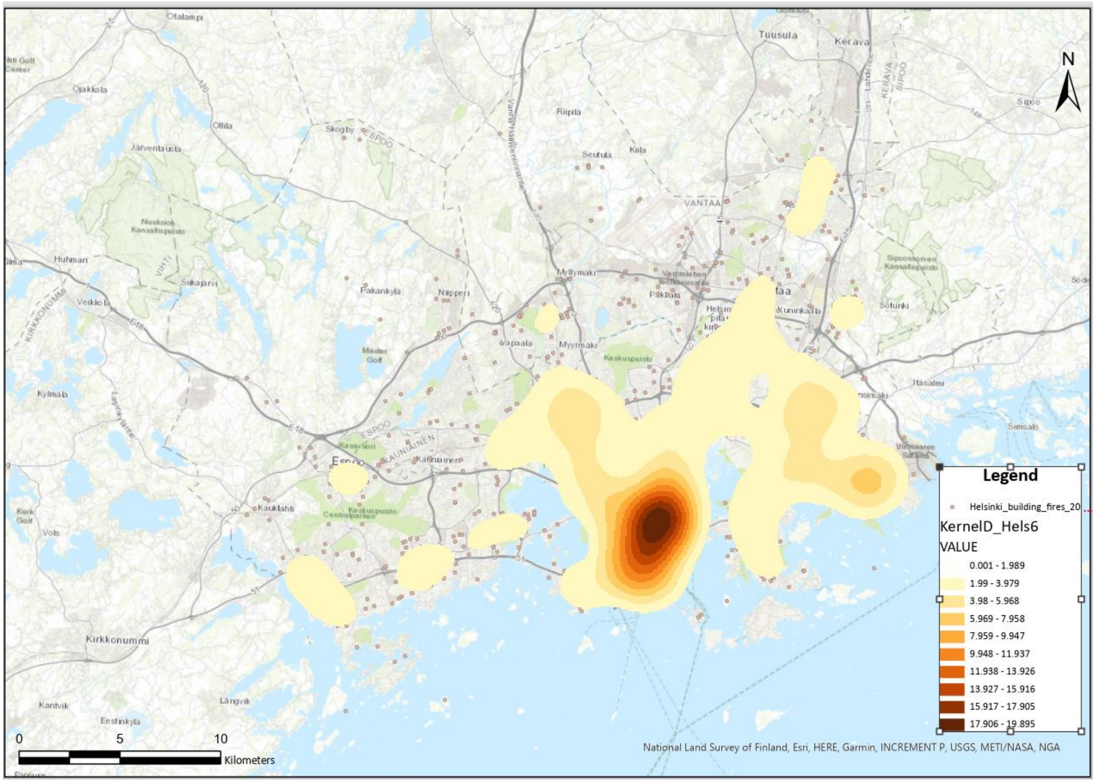

Fig 8 Kernel Density results with ‘Output cell size 70’ and ‘Search Radius 2500’

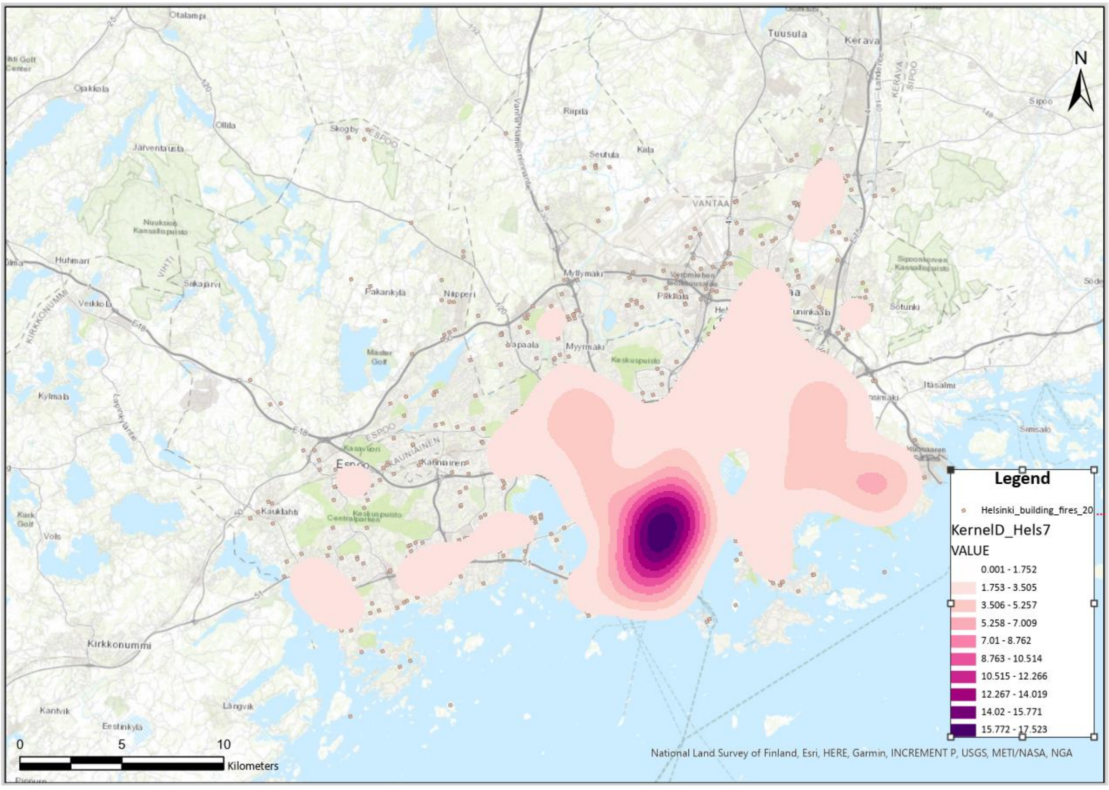

Fig 9 Kernel Density results with 'Output cell size 70' and 'Search Radius 2900'

Task 1.2

- **Question 1.1: Interpret the results of Ripley's K analysis you did. What does the results say about the input data? (1p)**

Q1.1: My answer：

The K-function analysis indicates that, across multiple distance bands, the observed K-values consistently exceed the expected K-values, clearly demonstrating that the building fires are spatially clustered rather than randomly dispersed. In other words, fire incidents are more likely to occur near one another, suggesting that certain areas of Helsinki experienced significantly higher concentrations of fires during the 2008-2010.

This clustering is observable at different spatial scales. By comparing the K-function charts for distance bands of 50 and 10, it becomes evident that clustering is present across various regions throughout the city. As the distance increases, the clustering remains evident, implying that these fires are part of a larger regional pattern rather than isolated local events.

From my perspective, this clustering can likely be attributed to several urban and environmental factors. High-density residential areas, older buildings with inadequate fire safety measures, or industrial zones with a higher risk of fire may all contribute to the observed spatial patterns. Overall, the Ripley's K analysis confirms that the Helsinki building fires from 2008 to 2010 were not randomly distributed but were spatially clustered, with significant concentrations observed at both local and broader scales.

- **Question 1.2: What does it mean if the ExpectedK is smaller than ObservedK? What about a case where ExpectedK is larger than ObservedK? (1p)**

Q1.2: My answer：

When the ExpectedK value is less than the ObservedK value, it suggests that clustering is present within the spatial data. The number of points observed at a specific distance exceeds what would be anticipated under a random spatial distribution, indicating that these points are situated closer together than would occur by chance, thus confirming the existence of clustering.

Conversely, when the ExpectedK value surpasses the ObservedK value, this signifies a dispersed state. In this scenario, the expected number of points at a given distance exceeds the observed count, implying that these points are more widely distributed than in a random arrangement and highlighting spatial dispersion.

By comparing ExpectedK with ObservedK values, one can assess and analyze the spatial structure of the data; higher ObservedK values indicate clustering while lower values suggest dispersion.

- **Question 1.3: What happens to Kernel Density analysis when Search radius is changed? What about if the output cell size is changed? (1p)**

Q1.3: My answer：

Search Radius:

The search radius is set to determine the distance around each feature (point or line), which helps in the calculation of density.

A larger radius includes more surrounding features in the calculation, resulting in a smoother, broader map of density. Density values will be less sensitive to small clusters or local variations, and areas with high and low densities will appear wider and more pervasive. This is useful for identifying broad trends or larger areas of concentration.

A smaller radius will create a more detailed and local density surface, focusing only on nearby features. This will result in a finer map, highlighting smaller clusters and providing more local detail. However, if the data has a fine scale, the results of many small hot spots may look complex.

Output Cell Size:

The output cell size determines the resolution of the resulting grid. It affects the detail or roughness of the final density map.

With a smaller cell size, the analysis output images will have higher resolution and can show more detailed density changes. This also means that each output unit covers a smaller area, providing a more accurate density value for the result. With a larger output cell, the density representation will be coarser, but will also capture broader density trends, but will lose fine detail.

- **Question 1.4: Look at the raster values of the Kernel Density analysis results you made with different search radii. Are the analysis results comparable or not? (2p)**

Q1.4: My answer:

As the search radius increases from 2100 to 2900, the areas of high density become larger and broader.

In Figure 7 (with a search radius of 2100), the density map shows more localized and smaller clusters of high density. The darker colors are more concentrated, indicating more distinct high-density areas.

In Figure 8 (with a search radius of 2500), the high-density areas begin to merge, with the dark regions expanding to a larger area, indicating a generalization of hotspots.

In Figure 9 (with a search radius of 2900), the high-density areas become smoother, with larger patches showing less variation within high-density regions.

As the search radius increases, the raster values in high-density areas generally decrease because the same number of points is distributed over a larger area. This effect dilutes the peak density values. Conversely, due to the larger radius encompassing more surrounding points, values in low-density areas may slightly increase.

The results are **qualitatively comparable** because they all analyze density using the same underlying data, allowing observation of trends in how the heatmap changes as the search radius increases. However, they are **not directly quantitatively comparable**, as the legends in each figure differ, reflecting different density values. The change in search radius affects the way density is calculated. A larger radius leads to a broader distribution of density values, lowering peak values.

Problem 2

Task 2.1

1. The Moran's Index and z-score values for the two datasets (1p)

   Answer:

   - The **Moran's Index** and **z-score values** for the Paavo postal code area dataset are **0.560402** and **11.803674**, respectively (fig 10).
   - The **Moran's Index** and **z-score values** for the Finnish population density dataset are **0.503162** and **25.770680**, respectively (fig 11).

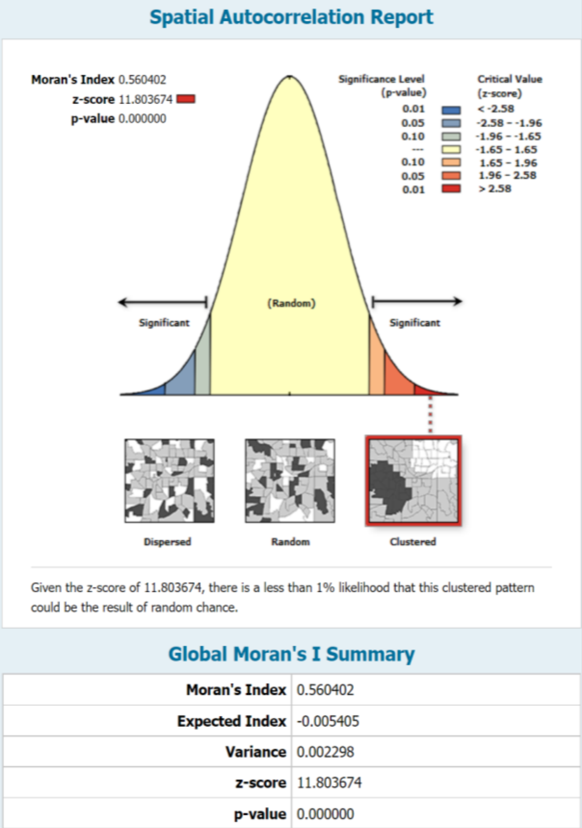

Fig 10 General report on Spatial Autocorrelation calculation for the Paavo postal code area dataset

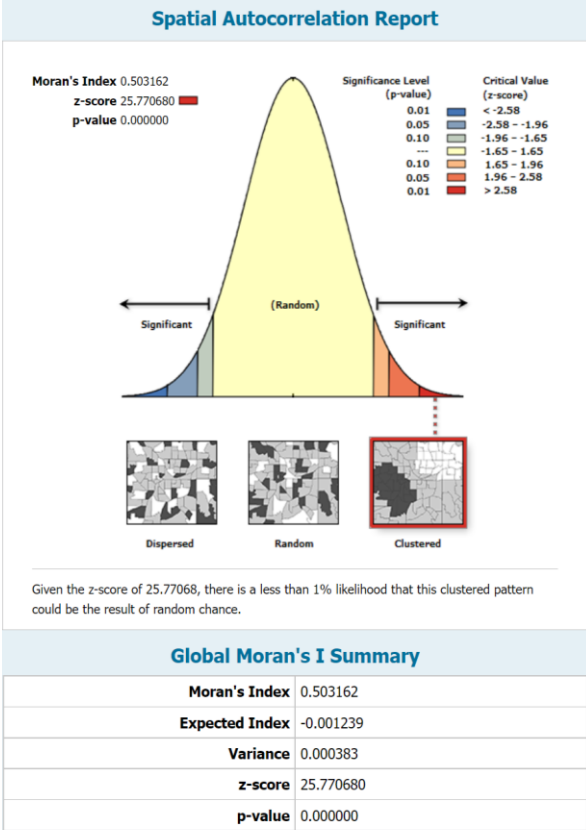

Fig 11 General report on Spatial Autocorrelation calculation for the Finnish population density dataset

2. Images of map views of the Cluster and Outlier Analysis results for the two datasets (1p)

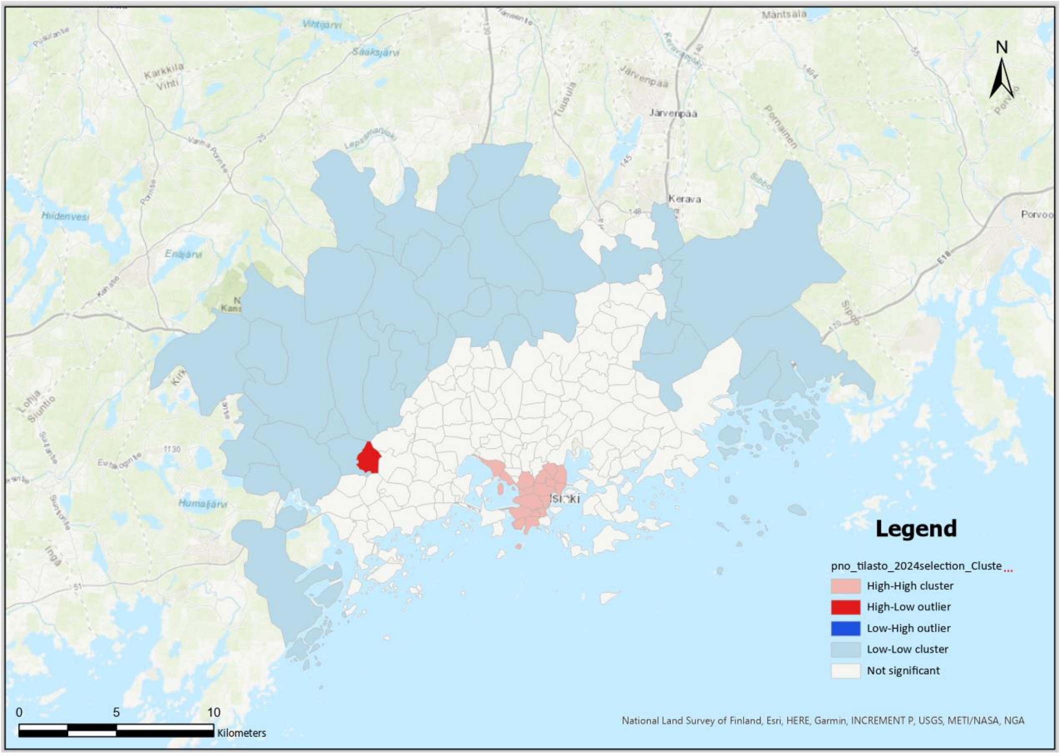

Fig 12 Map views of the Cluster and Outlier Analysis results for the Paavo postal code area dataset

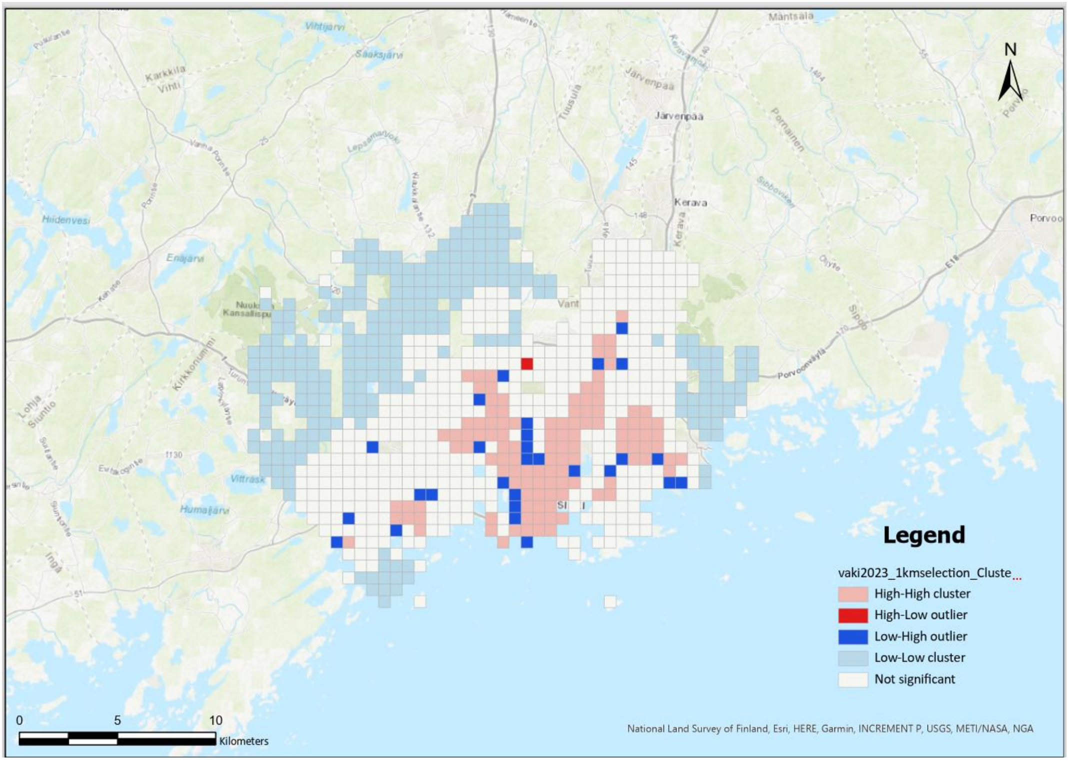

Fig 13 Map views of Cluster and Outlier Analysis results for the Finnish population density dataset

3. Images of Moran's Scatterplots for the two datasets (1p)

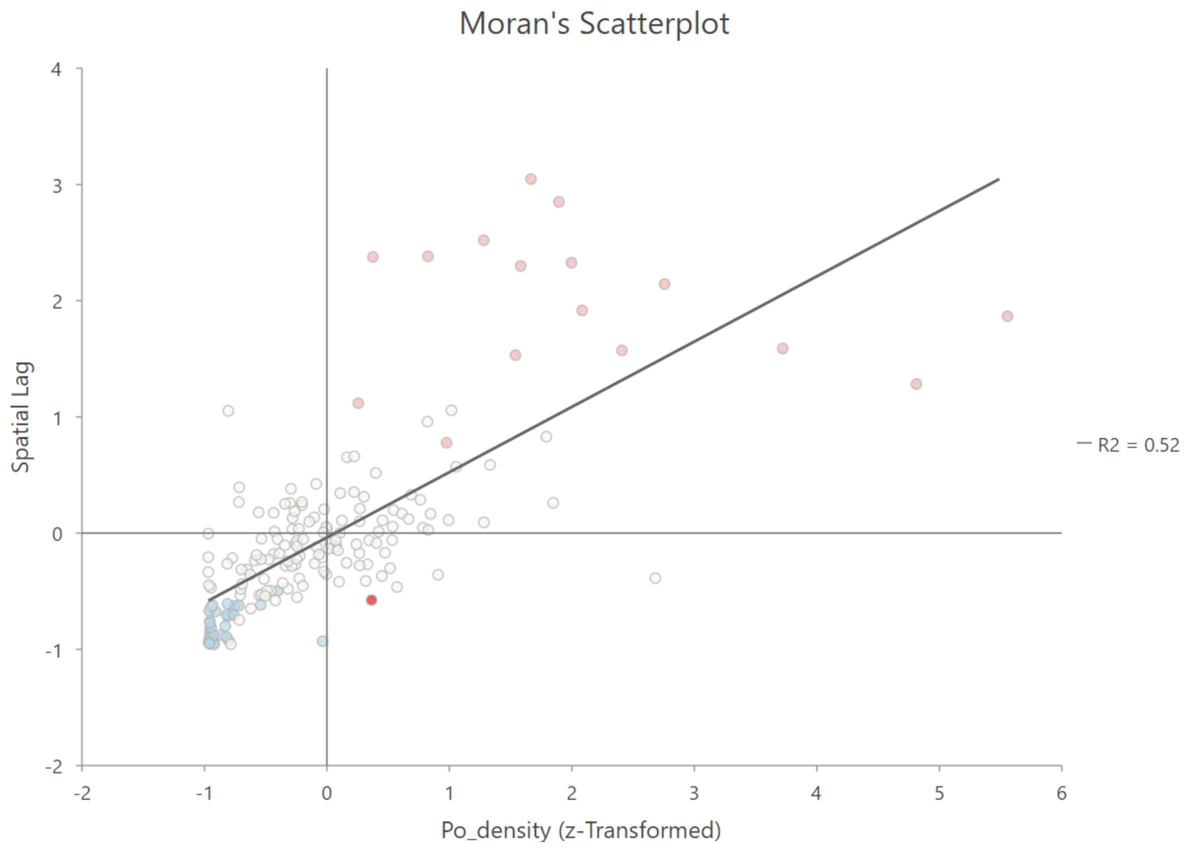

Fig 14 Moran's Scatterplots for the Paavo postal code area dataset

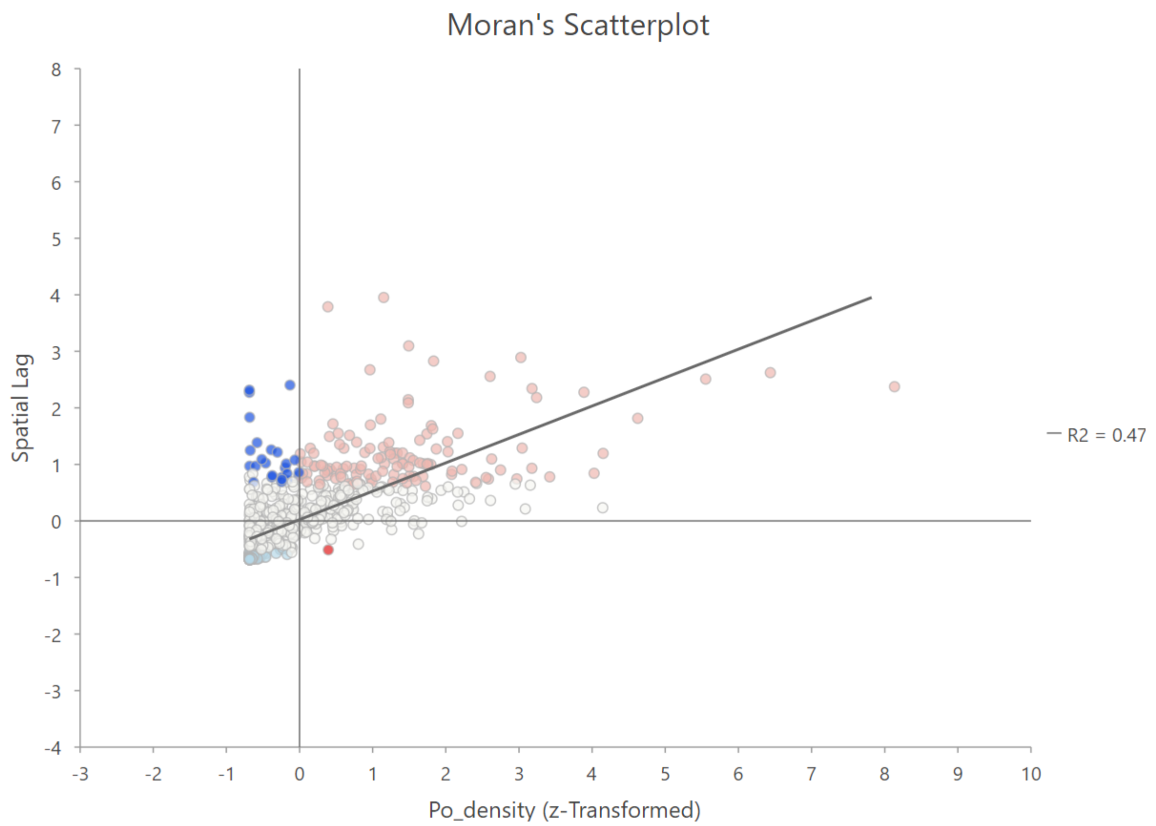

Fig 15 Moran's Scatterplots for the Finnish population density dataset

Task 2.2

- **Question 2.1: What do the Moran's Index values for the datasets mean? Is the data spatially clustered or not? (1p)**

Q2.1: My answer

Moran's Index is a statistical tool used to evaluate the degree of spatial autocorrelation within a dataset, indicating whether similar values are grouped together, spread apart, or distributed randomly across a region. By examining the specific Moran's Index values for a dataset, it's possible to identify whether the data follows a clear spatial pattern.

(1) A Positive Moran's Index (close to +1) suggests that similar values are clustered, meaning that areas with comparable data are situated near each other, forming distinct groupings.

(2) A Negative Moran's Index (close to -1) points to a dispersed pattern, where similar values are spaced apart, indicating that areas with comparable characteristics are not geographically close.

(3) When Moran's Index is around 0, it signifies a lack of spatial pattern, suggesting that the values are randomly distributed without noticeable clustering or dispersion.

- **Question 2.2: What do the attribute values LMiZScore RS and LMiPValue RS in the Cluster and Outlier Analysis results layer mean? (1p)**

Q2.2: My answer

LMiZScore RS is **a standardized Z-score** derived from the **local Moran's I**, which assesses how similar or different a location's values are compared to its neighbors. A high positive Z-score means that similar values are grouped together, indicating strong positive spatial autocorrelation. In contrast, a negative Z-score suggests that different values are clustered near each other, pointing to negative spatial autocorrelation.

LMiPValue RS is **the P-value** linked to the **local Moran's I**, reflecting the reliability of the spatial pattern observed. It shows whether the detected clustering or dispersion is statistically meaningful. A lower P-value, usually below 0.05, indicates that the spatial pattern is unlikely to be random, confirming the existence of a meaningful spatial relationship, either in clusters or in scattered patterns.

- **Question 2.3: What is the physical reason behind the high-low outlier polygon in the population grid dataset? (2p)**

Q2.3: My answer

This may result from a combination of city development strategies, economic factors, or environmental constraints that influence where people settle.

(1) Urban-Suburban planning:

High-low outliers are often found in areas that serve as a boundary between urban and rural regions. In these zones, you might have a highly populated urban cell right next to a less dense suburban or rural area. This sharp change is typical at city edges, where the population density drops significantly as you move further from the urban core.

(2) Economic and Infrastructure:

Some regions might show higher population density due to economic drivers, like job availability, closeness to major transit routes, or easy access to crucial infrastructure such as highways, schools, and commercial centers. These elements often create a dense population cluster within otherwise lower-density surroundings.

(3) Topographical/Environmental Features:

Geographical features can also play a significant role in creating outlier patterns. Natural elements like rivers, hills, or forests can restrict development in certain areas, leading to dense population clusters in more suitable locations while adjacent areas remain sparsely populated.

- **Question 2.4: What is the overall interpretation of the Cluster analysis results? Is this conclusion realistic? (2p)**

Q2.4: My answer

The Cluster Analysis results for both datasets (Paavo postal code areas and the Finnish population grid) indicate spatial clustering patterns in population density, revealing regions with high or low density adjacent to similar areas. In both datasets, areas with high population densities that are surrounded by similarly dense areas—known as high-high clusters—are mainly found in urban regions. Conversely, low-low clusters, where low-density areas are grouped together, are more prevalent in rural and suburban areas. The appearance of high-low and low-high outliers highlights transitional regions, where there's a sharp change in population density between urban and rural zones.

This clustering pattern is realistic given known urbanization trends in Finland, where population density is highest in cities and suburban areas and significantly lower in rural and northern regions. The cluster analysis results align with Finland's spatial development and urban planning policies, which concentrate population and infrastructure development in specific hubs.

- **Question 2.5: Are there significant differences between the Moran's I results for the density grid and the postal area dataset? (2p)**

Q2.5: My answer

There are differences in the Moran's I results between the two datasets. The Finnish population grid, with its finer resolution, tends to reveal a stronger spatial autocorrelation because it captures detailed changes in population density on a smaller scale. On the other hand, the Paavo postal code areas dataset, using larger spatial units, may show a lower Moran's I value.

These differences illustrate how scale and resolution can impact spatial analysis. A finer grid can pick up on subtle local differences, leading to more pronounced clustering patterns. In contrast, the larger units in the postal code dataset might result in a more generalized overview, showing less distinct clustering or dispersion patterns.

Problem 3

1. **To help us to develop the exercises, and understand the workload for you to complete the Exercise, please provide an estimate of how many hours you spent doing this exercise?**

   My answer: About 6 hours. For me, operating ArcGIS software is relatively straightforward compared to understanding the data, but I need to spend a significant amount of time (at least 4.5 hours) to grasp the meaning of Ripley's K, Kernel Density, Moran's Index, z-score values, Cluster and Outlier Analysis results, Moran's Scatterplots, and so on. I also need to understand the spatial characteristics these analyses represent, as well as the appropriate contexts for using these tools, in order to complete all the tasks. Despite this effort, I still feel that I am not able to accurately answer all the questions in the tasks.

2. **In addition, if you would like to give any feedback about the exercise, you can add comments under the Problem 3 (optional).**

   My answer: I find that the spatial data analysis content covered in Exercise 2 is quite extensive, making it challenging to apply and fully understand everything simultaneously. It would be more student-friendly, especially for those with a weaker background, if these theoretical concepts could be divided into two separate exercises. This way, students would have more time to digest and practice each analysis method, allowing for a deeper understanding of the material without feeling overwhelmed. Splitting the content would enable a more gradual learning curve, making the concepts easier to grasp and apply effectively.