Exercise 1

In Exercise 1 your task is to use desktop GIS software to solve two spatial analysis problems and provide information about the time it took to complete the exercise: Problem 1 (8 points), Problem 2 (11 points), Problem 3 (1 point).

The instructions for solving the exercise are given for the ArcGIS Pro software. We will not provide instructions or support for using other tools for solving the exercise. Also be advised that the analysis functionality available varies between different software. Therefore, should you wish to use e.g. QGIS to solve the exercise, your results might be a bit different.

Due date

Due date for this exercise is on **Friday, October 11th.**

Time allocation

We do not yet have a good estimate for the time it takes to complete this exercise as this is the first time we organize this exercise. However, a rough estimate is 3-8 hours.

What needs to be returned?

As an outcome of this exercise, you should write a short report that contains key visualizations of your work (screenshots are sufficient) and short descriptions of what they represent. In addition, the report should contain answers to all the questions that we ask related to the given problems/tasks. Please include intuitive headings to your report, such as “Problem 1 - Task 1” and indicate clearly which question you are answering, e.g. “Q1.1: My answer”.

Return the report into your personal Github repository as a single PDF file or a Word document. Name the file as “*exercise_1_MyGithubUserName.pdf*”

How to return the exercise to Github?

You can easily add a new file into your personal Github repository by:

1. Log into your Github account
2. Find and navigate to your personal exercise repository, such as *github.com/IntroSDA-2024/exercise-1-htenkanen*
3. Click “Add file” → “Upload files”
4. Drag and drop the report from your own computer into the repository
5. Add a short commit message (e.g. “Exercise 1 ready”)
6. Upload the file by clicking the “Commit changes” button

Hint: If you want to make changes to your report after you have submitted it to Github, you can simply upload the same (modified) file again using the **same filename**. Github will keep track of the changes to the file. Hence, you **don’t need to rename** the file when you have a different version of it.

Tools to be used

The analysis tasks in this exercise should be solved using a desktop GIS environment. We will provide you with instructions and support for solving the tasks using the ESRI ArcGIS Pro software.

Other tools are not needed for solving this exercise.

Hints about using the tools

Be advised that by default the ArcGIS Pro WFS API downloads a maximum of 3000 data elements from the API. This can be adjusted from the ArcGIS Pro Contents panel. Select the layer -> Properties -> WFS. Adjust the “Set the maximum features returned” value.

Introduction

In this exercise you practice the use of the ArcGIS Pro desktop GIS system for analyzing and visualizing spatial data in vector format.

In the first problem, you will familiarize yourself with the **Modifiable Areal Unit Problem** (MAUP) by comparing data visualizations from two different data sets that include population data. You will accomplish this by using the Paavo (postal number areas) and Väestöruutuaineisto (population grid) data sets to visualize the population density in the Helsinki capital region (Pääkaupunkiseutu). Your task is to extract the part of the data set that intersects with the Helsinki capital region, and visualize the population density in this part of the data set. The two visualizations need to be comparable, meaning they use the same classification and visualization.

In the second problem, you will assess the quality of two building data sets, the OpenStreetMap building data extract and the Helsinki Region Environmental Services HSY building data. The quality assessment will be done by comparing the data sets to the building data provided by the City of Helsinki. The City of Helsinki data is used as the ground truth due to the assumption that a large city like Helsinki has both the incentive and resources to maintain an accurate and up-to-date spatial data set about the buildings within the city area. In this exercise you will focus on the spatial quality characteristics (e.g. how well the data sets represent the actual buildings in the Helsinki area) of the data. Attribute data quality is not included in the analysis.

Input data

Problem 1

You can find all the required data in the Paituli service. For the first problem we suggest you download the required datasets from Paituli in order for the work to go smoothly. Optionally, you can also use the WFS API for the data, but this may be a cumbersome approach with the Population Grid due to the amount of elements in the dataset.

Paituli Download page: https://paituli.csc.fi/download.html
Use the newest available versions of the **Paavo** (open data postal code areas) and

Population grid data (open data 1km x 1km population grid), both by Statistics Finland. In addition, you will need the **Administrative borders** dataset by National Land Survey.

Paavo data description can be found here:
https://stat.fi/tup/paavo/paavon_aineistokuvaukset_en.html

The population grid data description can be found here:
https://www.paikkatietohakemisto.fi/geonetwork/srv/eng/catalog.search#/metadata/a901d40a-8a6b-4678-814c-79d2e2ab130c

Problem 2

For the second problem you need to fetch the data from different sources. First, download the latest version of the OpenStreetMap data extract maintained by Geofabrik in ESRI Shapefile format. The data is updated daily, and thus represents the most up-to-date data available in OSM. You can find the data from here: https://download.geofabrik.de/europe/finland.html

Note that the Geofabrik data extract covers the whole country of Finland.
The other two data sets you can access via the appropriate API.

City of Helsinki WFS API:
https://www.hel.fi/en/decision-making/information-on-helsinki/maps-and-geospatial-data/make-better-use-of-geospatial-data/open-geographic-data#open-data-interface-service
Layer **rakennukset_alue**

HSY WFS API:
https://www.hsy.fi/en/environmental-information/geographic-information/geographic-information-interfaces/open-geographic-information-interfaces/
Layer **pks_rakennukset_paivittuva**

Note that there's less than 200 000 buildings in the Helsinki capital region. Adjust the WFS API settings in order to fetch all the data elements (see a hint about this above). Be advised that it can take a moment to download all the data. By default ArcGIS Pro caches the data from the API, so after the initial download the data should work smoothly and without data transfer lag.

ArcGIS functionality used

In order to solve the problems in this exercise, you need at least the following ArcGIS Pro functionality.

Adding data and WFS connections. Data on your local machine can be added to an ArcGIS map by just drag&dropping. For WFS Connections, you need to either use the Catalog panel (Servers->Add Server) or the Insert tab from the Ribbon top-level UI (Insert->Connections->Server->New WFS Server). Remember to have the URL of the server to add available.

Selection and creation of new layers from selected features. The Select by mouse cursor and Select By Location -functionality are useful in the exercise. You can turn selected features into a new Layer by right-clicking on the layer in the Contents panel and selecting either

Data->Export Features to create a new data set from the selected features or Selection -> Make Layer From the Selected Features to directly create a new layer in the map. For Problem 1 it is sufficient to create new layers from the selected features, but in Problem 2 you need to Export some data sets.

Symbology. You can adjust the symbology of a layer by right-clicking on the Layer in the Contents panel and selecting Symbology. This opens the Symbology panel where you can adjust the way the data is visualized.

Repair Geometry. The Repair Geometry tool can be used to filter out broken data elements from a data layer (e.g. non-simple data elements, broken polygon boundaries, degenerate data elements, etc.). In ArcGIS Pro many tools expect the input data to be valid and cannot work with broken data elements.

Erase. Calculates the difference between elements of the the first and second input layer (Input Features and Erase Features in ArcGIS Pro terminology). Can be used to calculate which parts of the first layer are not present in the second layer.

Problem 0 - Slack assignment

This part is only for registered students at Aalto. We have sent an invite link to Slack to all registered participants before the first lesson. Please contact the course instructors via email in case you are registered for the course, but don't have access to Slack. In this problem, you should report your Github username to us via Slack (needed for grading).

Steps:

1. Go to our Slack page at https://introsda.slack.com
2. Post a new message in the **#exercise-1** channel with your **full name** and **GitHub username**
    - For example: `Henrikki Tenkanen,htenkanen`

Problem 1 (8 points)

In this problem the goal is to visualize the population density in the Helsinki City Region using two datasets: the postal code areas and the population grid. The data visualizations created in this exercise must be comparable. This means that with both datasets, the same visualization method needs to be used in the same way. In this exercise, it is probably the best if you classify the data into 5-7 classes. Remember to make the classification comparable, meaning that visualizations for the two data sets represent the same phenomenon in the same way. For example, if you use classification, use classification in the same way for both datasets.

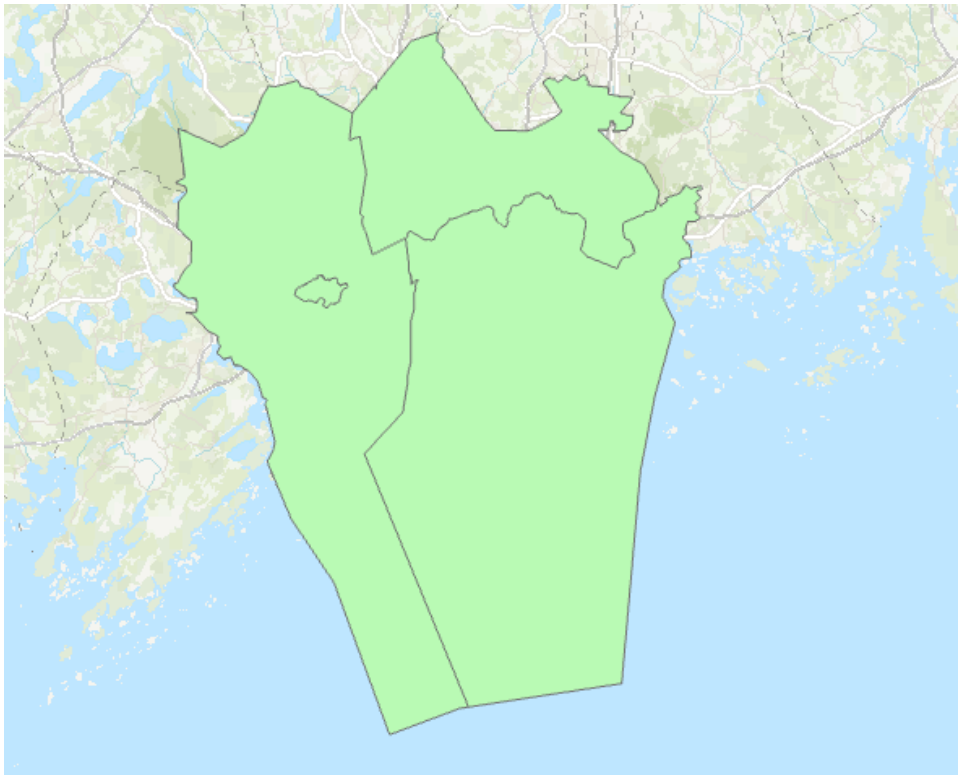
The ArcGIS Pro tools needed for solving this problem are data selection and symbology manipulation.

Task 1.1

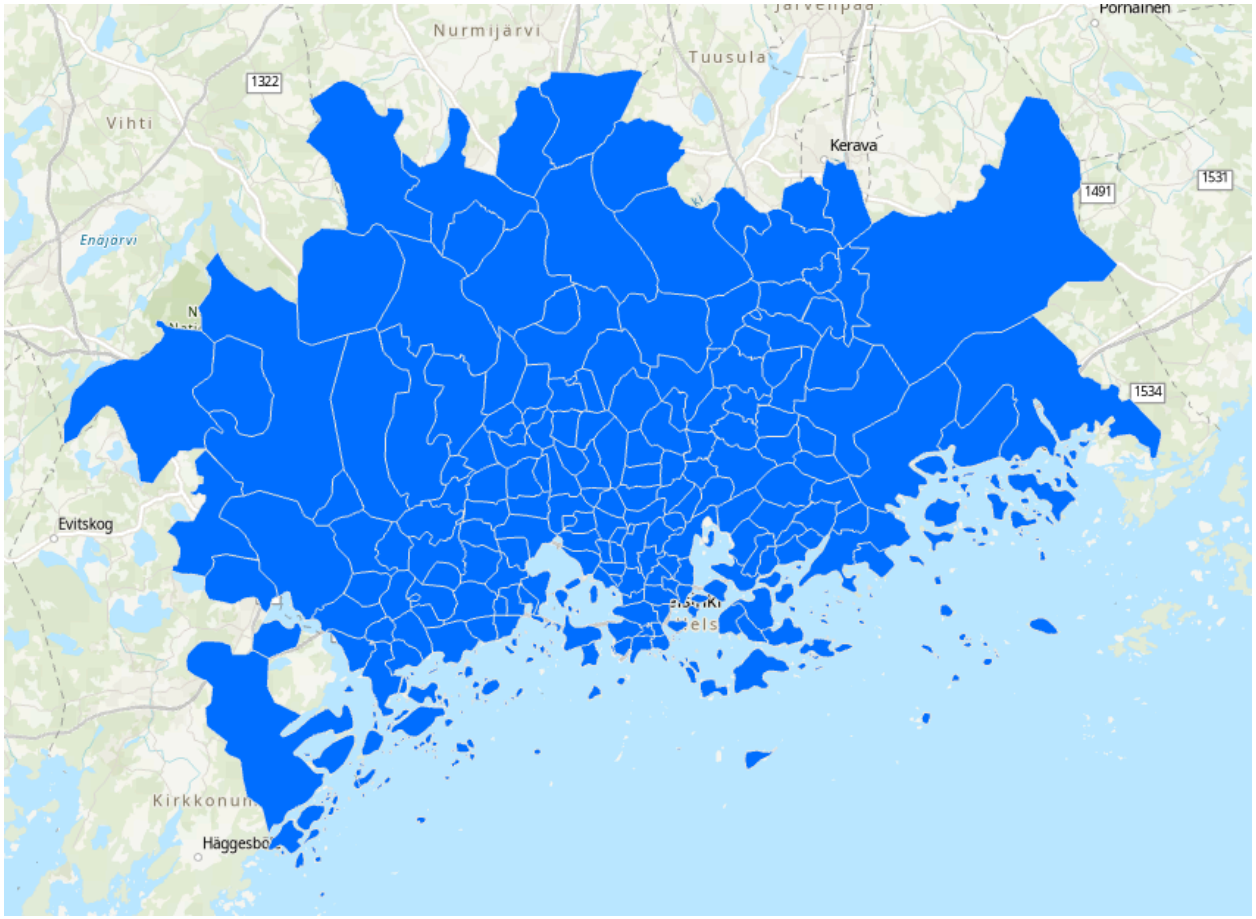
In task 1.1 you create the appropriate population density visualizations for the two datasets.

To solve task 1.1 Do the following:

- Fetch data required for Problem 1.
- From the municipality data, create a new layer containing the polygons for the Helsinki Capital Region (the cities of Helsinki, Espoo, Vantaa, and Kauniainen). See image.



- Use the Capital Region polygons to select the postal areas that **intersect** any of the capital region polygons. Do the same for the population grid polygons. See also the image below, which shows all postal number areas that intersect the Helsinki Capital Region.



- Visualize the newly created layers with appropriate visualizations that show the **population density per square kilometer**. Note that the population grid is already in square km size, but for the postal areas you need to adjust the visualization appropriately.
    - In ArcGIS the population density calculation can be done in the visualization tool
- Adjust the visualizations so they are comparable (e.g. number of classes and division of data into classes). We suggest using natural breaks for classification and to use 5 classes.

In your answer provide the following:

1. **Image of the population density using the postal code areas** (1p)
2. **Image of the population density using the population grid** (1p)
3. **Image of the map legend for both the postal code and population grid representations of the population density (1p)**

Task 1.2

In task 1.2 you will compare the two visualizations and answer the following questions. Write a few sentences about your thoughts as an answer for each question:

- **Question 1.1:** Explain how you have ensured that the representations of population density in the two maps are comparable? (2p)
- **Question 1.2:** What are the most important similarities and differences between the representation of population density on the two maps? (2p)

- **Question 1.3:** Is this a good example of the Modifiable Areal Unit Problem? If so, why? (1p)

Problem 2 (11 points)

In this problem your task is to compare the Helsinki Region Environmental Services (HSY) and OpenStreetMap (OSM) building datasets against the City of Helsinki (HKI) building dataset. This is done to evaluate the spatial quality of HSY and OSM building datasets. The underlying assumption in this work is that the City of Helsinki maintains the most spatially accurate and up-to-date dataset on the buildings within the city. Therefore the HKI dataset is used as a validation dataset to measure the external quality of the HSY and OSM datasets. The quality evaluation in this exercise is limited to spatial data quality, meaning **how well a dataset corresponds to the real world situation it represents**.

In more detail, we're interested in **spatial accuracy,** meaning how well data elements correspond to the locations of the real world elements they represent, as well as **completeness**, meaning are all real world buildings represented in the data. For completeness, we're interested both in missing data (real-world building not included in the data set) and extra data (data element not corresponding to real-world building in the data).

The tools required for solving this exercise are ArcGIS Pro data selection, data exporting, geometry repair, and erase.

Task 2.1

In task 2.1 you will create new datasets that represent the difference between HKI and HSY as well as HKI and OSM datasets.

Hint: Create a new map in the ArcGIS Pro project for this problem. It is easier for you to keep track of the appropriate data that way.

To solve the task 2.1, do the following:

- Fetch the data. Remember to adjust the WFS settings in order to get the full datasets for City of Helsinki and Helsinki Region Environmental Services. For visualization, below is the Katajanokka neighborhood with buildings in the HSY dataset.

- For OSM and HSY data, create subsets that contains all data elements within the area of Helsinki for further processing. Save these subsets locally for further processing.
- Create a local copy of the Helsinki dataset for further processing.
- Some of the data sets may contain broken geometries. Repair the data sets using the Repair Geometry -tool. Repair the data to follow the OGC Simple Features standard.
- Use the Erase tool to compare the HKI dataset to the OSM and HSY data.
- Save the comparison results in a format that is easy for you to use for further analysis

In your answer provide the following:

1. **Image or images clearly showing the differences between HKI and HSY data for the neighborhood of Katajanokka in Helsinki (https://en.wikipedia.org/wiki/Katajanokka) (2p)**
    a. Make sure it is easy for the reader to see the locations where there are differences
    b. Include an explanation of exactly what the image(s) represent(s)
2. **Image or images clearly showing the differences between HKI and OSM data for the neighborhood of Katajanokka in Helsinki (https://en.wikipedia.org/wiki/Katajanokka) (2p)**
    a. Make sure it is easy for the reader to see the locations where there are differences
    b. Include an explanation of exactly what the image(s) represent(s)

Task 2.2

Assess the results of your analysis and answer to the following questions regarding the quality of HKI, HSY and OSM data sets

- **Question 2.1:** When compared to HKI data, what is your assessment of the spatial accuracy and completeness of HSY data? What are the biggest problems in the quality of the dataset? Justify your answer. (2p)

- **Question 2.2:** When compared to HKI data, what is your assessment of the spatial accuracy and completeness of OSM data? What are the biggest problems in the quality of the dataset? Justify your answer. (2p)
- **Question 2.3:** Which data set is of higher quality, OSM or HSY? Why? (1 p)
- **Question 2.4:** Was it a reasonable assumption to use HKI data as a baseline for external data quality assessment? Justify your answer. (2p)

Problem 3 (1 point)

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?**

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