Technical Assessment: SEO

Background

You are a modeller in a health protection team in the UKSHA.

Imagine there has been an outbreak of an airborne pathogen from an unknown source. All that is currently known is the locations of cases at the time of the release. The Outbreak Control Team has come to you to estimate epicentre of the outbreak and the number of people who may live within the (currently unknown) affected area.

Aim

The aim of this assessment is to see how you handle a representative modelling task. We do not expect all candidates to finish the assessment, we are far more interested in your approach to the problem. The assessment is used to guide a discussion about how you would handle similar tasks.

In this assessment, we are looking for evidence of the ability to:

- join appropriate methods together to solve problems,
- take an algometric description and either turn it into useful, reusable code, or locate an appropriate existing implementation,
- to work with data and mechanistic models to provide solutions to problems,
- to visualise data and model outputs so that decision makers may act on your work.

During the panel session, you will be asked questions about:

- how you approached the task,
- your findings and conclusions,
- understanding of approaches used,
- further ideas you would have explored/implemented with more time

You will be allocated **90 mins** for this assessment. At the end of the 90 minutes please push your code and / or notebooks to the Github repository you created for this task and add Github users "twomagpi" and "SDykeUKHSA" to that repository with at least read access.

Full Task Description

Your manager has sent you two files, you will likely need both to answer the question:

1. A case line list dataset – **2023-10-15_Modelling_test_case_locations_v1.0.csv** they have already converted the address locations into X and Y co-ordinates in the same co-ordinate reference system as the population grid. A data dictionary is attached at the end of this document.
2. A matrix containing a gridded representation of the population numbers for the area concerned – **2023-10-15_Modelling_test_population_v1.0.csv**
    a. This is a geographic dataset so 0,0 is in the bottom, left corner of the matrix.
    b. Each cell is 100m east-west by 100m north-south
    c. The numbers in each cell represent the number of people who were assumed to be in that cell at the time that the pathogen was released. For the purposes of this test you may assume that they were all at the centre of the cell at the time of the release. (So for example the 42 people in the bottom left cell may be assumed to have been at 50,50.)

Outputs

The outbreak control team would like to understand where the centre of current outbreak is, and what the possible affected population numbers are.

Your manager would like you to demonstrate clear steps in how you approach this problem. The code you create and that you understand the algorithms you used.

You may assume that the pathogen has spread no further than the furthest cases, and that it spreads evenly in all directions – i.e. forms a circle. You may wish to consider this part of the test as an example of the minimum bounding circle problem.

When you have determined the extent of the spread, you will then need to count the population affected.

Your manager and the outbreak team would like to see at least one appropriate visualisation using a plotting library of your choice.

Tools

You are free to work with any analytical environment that you have available to you. We would prefer candidates use appropriate open-source data science tools, such as Python or R. If you cannot install Python or R locally, please use Google Colab (https://colab.research.google.com/) and save as a private notebook – this requires a Google account.

You may use existing public libraries (packages) but please expect to be asked about the underlying algorithm they implement and why it is appropriate.

Please do not share this material with anyone else, and do not upload this dataset, or code and outputs you produce to a public repository (e.g. GitHub, Kaggle).

Data Dictionary

Field	Description	Type
id	Identification of a patient	Integer
X	The X (easting) co-ordinate, in meters, of the patient's location at the time of the release	Float
Y	The Y (northing) co-ordinate, in meters, of the patient's location at the time of the release	Float