

# Challenge 1: Spot the outliers

## Challenge Description

The Hollandse Luchten project makes use of SODAQ Air sensors to measure air quality, in particular particulate matter (PM2.5). Due to several factors, these sensors can produce faulty or anomalous measurements. Such measurements can have a big impact on downstream analysis, other tools that use the data and/or other follow-up actions that are taken. Therefore, in this challenge your task is to work on an automated approach to 'spot these outliers'.

## Problem Statement

By spotting an outlier, we identify a data point that is distinctly different from other data points. An unusual measurement is thereby not automatically an incorrect measurement. A few examples of why outliers may occur:

- Sensor calibration drift over time,
- Electronic faults or data transmission errors,
- Environmental/weather conditions,
- New years firework,
- Unusual but valid atmospheric conditions and
- Other local hotspots / sources of pollution.

An important element in identifying outliers, might be the context in which you assess the data point and how certain are we that the data point is an outlier? Is the datapoint an outlier *within* the data from that specific sensor, in comparison to other Hollandse Luchten sensors in the region and/or compared against official reference measurements?

Therefore, in this challenge you will focus on:

1. Using data science and AI techniques to automatically **detect** outliers.
2. A scoring metric to **indicate** the status of the sensor at a moment in time and **how certain** the techniques are about the outlier.
3. **Communicating** this scoring metric to residents in a clear and transparent manner and **why** it is considered an outlier.
4. **Bonus**: indicating to the residents for what kind of analysis the sensor in that time period can be used for and for what kind of questions it cannot. And possible follow-up actions that need to be taken in case of a malfunctioning sensor.

## Possible techniques

Teams may use statistical, machine-learning, deep learning and/or hybrid methods. A simple and well-validated approach may be more valuable than a complex model that is difficult to explain or verify. We expect from teams that they are able to substantiate (validation metrics, research, others) any made technical choices and considerations to the jury.

## Which data is available?

During the hackathon we will supply you with several datasets for this challenge:

- Time-series data: hourly PM2.5 and PM10 from more than 100 Hollandse Luchten sensors, some ranging several years.
- Reference measurements: Official reference station data producing hourly measurements, to possibly compare the Hollandse Luchten sensors with.
- Metadata on sensors: information about the sensor location, type, measurement communities, and intended reference stations.
- Weather data from the Dutch KNMI, for example temperature and humidity.
- Historical quality information: previously identified faulty sensors, known artifacts, and periods of degraded quality.

## Deliverables

Your submission should at least contain:

- 1) The Github repository with any used code and/or examples.
- 2) The presentation of your solution.

Every team is expected to present their solution to our jury.

## Jury criteria

*Will be added later*

## Use of tools and generative AI

AI (coding) assistants and vibe coding are allowed, as long as teams disclose clearly on where they did and did not use them. We encourage the use of such tools, but also address that teams should be able to explain and validate important parts of their solutions themselves.

## Questions?

The full repository will be shared a week before the hackathon. If you have any questions, please e-mail them to [janus.jansen@noord-holland.nl](mailto:janus.jansen@noord-holland.nl). These and spontaneous questions will be answered during our 'Online Q&A' on 6 November 15:00-16:00. Participants will receive the Q&A meeting link via email.