Imperial College London

High Energy Physics Group

Supervisor : Mark RICHARDS

Characterization of pollution events in simulated sensor network datasets

Cephren REES

Background

- Air pollution is a danger to public health,being the leading environmental risk factor for mortality/ ill health. (Air pollution and health | UNECE, 2021)
- Monitoring pollutant levels can assist community response to the challenges posed by this threat.
- Sensor networks represent a valuable new paradigm in this regard.(Ma et al., 2008)
- Characterization of pollution events from network generated datasets develops critical analytical tools and ensures a set of benchmarks for fidelity and accuracy when transitioning to a roving ad-hoc sensor network.

Research Aims

- Visualize evolution of pollutant levels from raw datasets.
- Develop protocols for identifying pollution events.
- Thoroughly characterize events based on salient features.

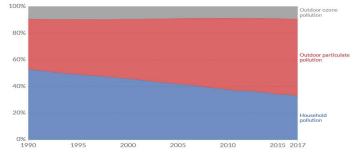

Figure 0: Death rates from air pollution, World, 1990 to 2017, relative(Death rates from air pollution, 2021)

Methodology 1.a

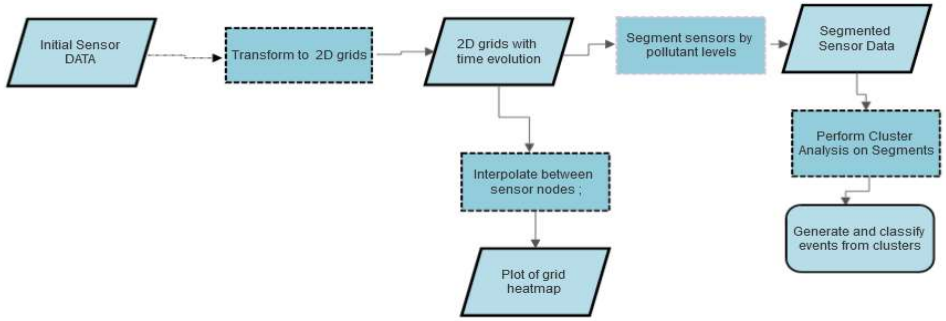

Figure 1: Flowchart describing investigation process.

Methodology 1.b

- Splines (piecewise defined polynomials) were used to interpolate between distinct points in order to construct a heatmap with significant resolution.
- Various cluster analysis methods were tried but the most appropriate was found to be density-based clustering. (Sander, J. et. al.,1998)
- Consistent parametrization throughout the datasets is better implemented with this method and it accounts well for outliers and arbitrary cluster shapes.
- The method is based on principle of ϵ-neighbourhood for points p, q of dataset such that

$$N_{\varepsilon}(p) : \{q \mid d(p,q) \leq \varepsilon\}$$

- Rates of information loss for clusters throughout timesheets define duration of event.

Results

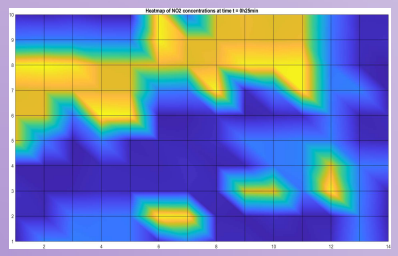

Figure 2: Heatmap with visible pollution events with respect to background, 'hot' regions are snapshots of events

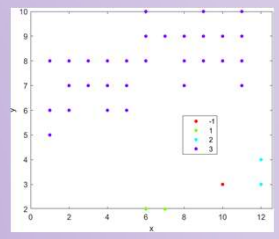

Figure 3: Corresponding Cluster Breakdown of Fig.2

- Primary protocol of event detection Is cluster analysis.
- Density based clustering accounts for arbitrary shapes and outliers, data segmentation and initial parametrization crucial to algorithm implementation.
- Ongoing analysis of cluster features.
- Interlinking clusters throughout timesheets into events is key to full event characterization.

Conclusions

- Density based clustering ideal for uniform parametrization
- Quantifying information loss will be key to defining extent of an event
- Revisions to segmentation and clustering algorithms expected.
- Future focus on grouping clusters through time into events
- Not all clusters are equal when defining events.

References

- Unece.org. 2021. *Air pollution and health | UNECE*. [online] Available at: <https://unece.org/air-pollution-and-health> [Accessed 3 March 2021].
- Ma, Y., Richards, M., Ghanem, M., Guo, Y. and Hassard, J., 2008. Air Pollution Monitoring and Mining Based on Sensor Grid in London. *Sensors*, 8(6), pp.3601-3623.
- Our World in Data. 2021. *Death rates from air pollution*. [online] Available at: <https://ourworldindata.org/grapher/death-rate-by-source-from-air-pollution> [Accessed 3 March 2021].
- Sander, J., Ester, M., Kriegel, H.-P., Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery*, *2*(2), 169–194.