24th INTERNATIONAL CONGRESS ON SOUND AND VIBRATION
23–27 July 2017, London

EXPLORING THE 'BIG ACOUSTIC DATA' GENERATED BY AN ACOUSTIC SENSOR NETWORK DEPLOYED AT A CROSSRAIL CONSTRUCTION SITE

Ben Piper and Richard Barham

Acoustic Sensor Networks Limited,
email: ben.piper@acousticsensornetworks.co.uk

Steven Sheridan

Crossrail Limited,

Kostas Sotirakopoulos

National Physical Laboratory

A distributed network of acoustic sensors allows for the spatial properties of sound in an environment to be measured, giving information that is not available when measuring in a small number of locations. This type of sensor network has only recently become technically feasible and economically viable due to developments in low-cost sensors and computing hardware. The use of a large number of sensors generates big and rich datasets which can be problematic to analyse and visualise using traditional approaches. There are a number of machine learning tools which can be applied to this type of dataset and the use of carefully designed data visualisations allow for the stories in the data to be told.

In this paper data measured at a Crossrail construction site using a network of 16 sensors will be analysed to show new insights that can be found when using an acoustic sensor network, focusing on the spatial properties of the sound field measured and the performance of the individual sensors and the network as a whole.

Keywords: Data, Sensor-Network, Smart-Cities, Environmental Noise

1. Introduction

Measuring and controlling the noise produced by a construction site in an urban environment is a difficult task due to the complexities of the sound generated by the site and its surrounding area, especially from neighbouring construction sites. In the UK, planning controls are set using Section 61 of the Control of Pollution Act [1]. This is not a standard framework, with new constructions sites mandated to consider the conditions of existing construction sites in the locality. This is essentially a 'first come first served' approach to setting noise limits driven by the difficulty in assessing the cumulative noise effects of neighbouring sites. To separate the noise generated by a single site from its surroundings a new approach to measurement is required that allows for contributions to be attributed. This in turn can lead to a more standardised approach to noise planning controls that doesn't penalise one development for the existence of another. Achieving this requires measurements of the spatial distribution of the noise which, until recently, was prohibitively expensive and difficult to implement.

With advances in technology such as MEMS microphones, mini-computers and big data analytics these challenges can now be overcome. Several studies have explored the use of consumer grade electronics for creating distributed networks of acoustic sensors [2-4] with a number of different applications such as smart traffic management [5], quantifiable soundscape research [6] and the health effects of noise exposure [7].

In 2016, a large measurement campaign was conducted at Crossrail's Moorgate construction site in London, UK, using noise signature correlation techniques to successfully separate the noise generated by the site from its surroundings [8]. This study proved very successful and the techniques developed within it are readily scalable to multiple construction site scenarios.

This paper will explore the dataset generated by this campaign for additional information, making use of processing and visualisation tools designed for handling large datasets. It will not repeat the correlation analysis work undertaken during the project but will focus on using measurement statistics to find patterns in the data. The aim here is to demonstrate approaches which can offer easy ways to visualise the features of a large noise measurement dataset, focussing on the availability of the data and the temporal and spatial distribution of the noise levels.

2. The measurement campaign

The measurement campaign at Crossrail's Moorgate site consisted of 16 measurement nodes deployed on the edges of a construction site with the aim of demonstrating a method for separating the noise contribution of the construction site from noise generated by the surrounding environment including the neighbouring Crossrail construction site. The study focussed on the use of noise signature correlation techniques for identifying the source locations.

2.1 The measurement nodes

The measurement nodes consisted of a MEMS microphone package and a Raspberry Pi 2B mini-computer with audio and communication peripherals. The computer was housed inside a water and dust proof box (IP65) with the microphone connected by a short cable and held at a distance by an aluminium tube. The systems where powered by 5V DC power supplies connected to the construction site's mains supply.

The microphones consisted of a MEMS microphone housed inside a 7mm diameter stainless steel tube and fitted with a patented acoustic filter [9]. The filter corrects the microphone's frequency response for the resonance of the MEMS package and the diffraction caused by the tubular housing whilst offering a level of protection to the microphone from dust, moisture and wind. The microphone is also fitted with a hydrophobic windshield to further protect it from the environment and to limit the amount of wind noise contaminating the measurements.

The computers processed the signals, calculating broadband and third octave $L_{A,eq}$ levels every 0.2 seconds. This data was transmitted to a database over the 3G/4G network where further software was used to aggregate and present the data via a web portal.

2.2 The layout of the site

The test site for the measurements was the west half of the Crossrail development at Moorgate in central London, UK, shown in Fig. 2. The east half of the site contains the main shaft of the station. For the first half of the measurement campaign this site was managed by a different construction company from the test site. Approximately 100 meters to the west of the construction site is Willoughby house which is a 6 storey 148 flat terrace block on the eastern edge of the Barbican complex and is the main sensitive neighbour. To the east of the site is the busy Moorgate thoroughfare and further to the east is Finsbury Circus and Liverpool Street Station.

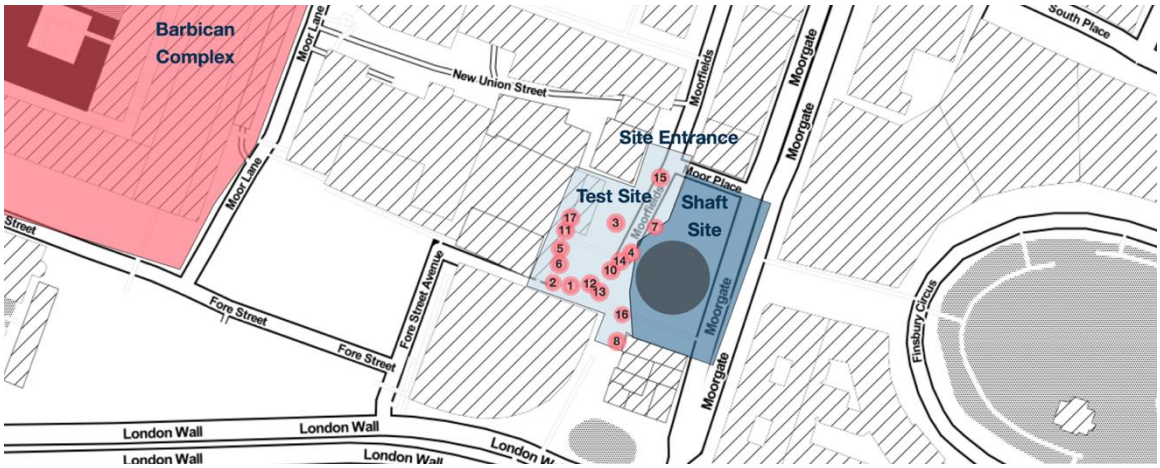

Figure 2: Map of the construction site and neighbouring features with node positions marked.

A ring of 14 nodes was placed around the test site with 2 further nodes positioned on the north (node 15) and south (node 8) extremities of the site. The delivery entrance to the site was on the north side of the site. The site featured 3 levels with nodes distributed on all 3. nodes 17, 11 and 5 were situated on the upper level in the accommodation area of the site. nodes 16, 14 and 4 were situated on the lower level on the eastern edge of the site overlooking the shaft. The remaining 8 nodes were positioned around the edges on the middle level with nodes 3 and 7 on the north side near the delivery entrance.

2.3 The datasets

One day of the raw dataset is approximately 1.3 GBs in size giving a total dataset size of over 250 GBs. For this paper 27 weeks of aggregated 1 minute data and 4 one day subsets of the raw dataset and have been used. The data was cleaned and fitted into fixed size data frames. The details of the datasets are given in Table 1. These datasets can be considered 'quite big' data but the type of measurements made are very scalable and the techniques applied to big data can be employed here.

Dataset	Start Date	Duration, days	Resolution, seconds	Size, GB	Size, rows	Size, columns	Size, cells
1	27/04/2016	189	60	0.76	4,354,560	29	126,282,240
2	24/06/2016	1	0.2	1.51	6,912,016	30	207,360,480
3	23/07/2016	1	0.2	1.51	6,912,016	30	207,360,480
4	18/08/2016	1	0.2	1.51	6,912,016	30	207,360,480
5	20/08/2016	1	0.2	1.51	6,912,016	30	207,360,480

Table 1: Details of the datasets used for the analysis in this paper.

3. 27 weeks of 1 minute data

In this Section analysis of the data availability and level distributions of Dataset 1 will be shown

3.1 Data Availability

A useful first step in understanding a dataset is to analyse its completeness. The dataset is fitted to an expected time array with null values placed where there is no measured data. Using the Python Missingno library, this is visualised as an availability matrix in Fig. 3. For the timescales shown, the smallest visible gap represents a period of between 1 and 6 hours where most data are missing. This highlights several patterns in the availability of data; Nodes 5, 11 and 17 record no data for the last 8 weeks of the campaign and for 2-3 days after the 19th of June; Node 8 did not record any data until 7 weeks into the campaign; All the nodes are missing data for approximately 1 day in mid-July; Node 15 has intermittent availability for the final third of the campaign.

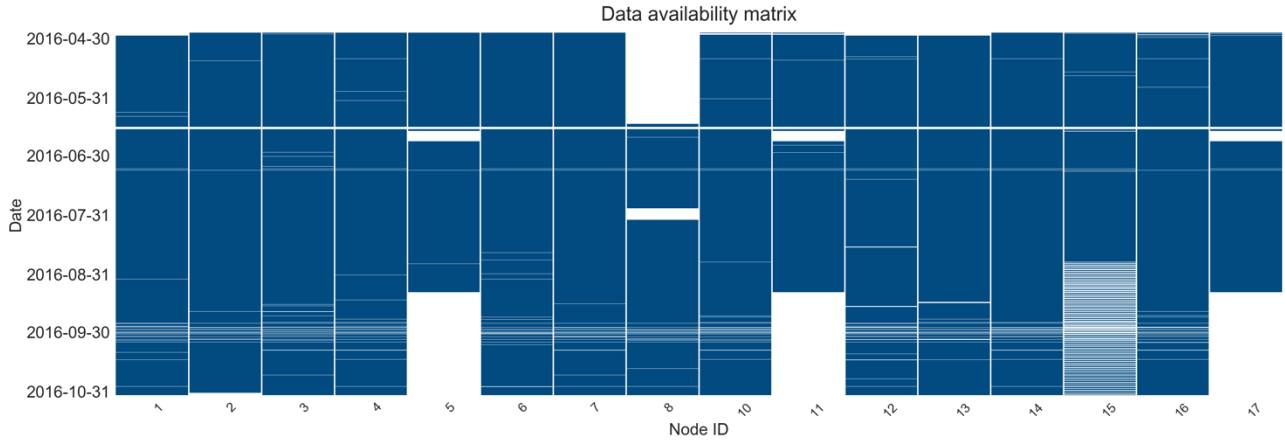

Figure 3: Data availability matrix for all 16 nodes between the 27th of April and the 2nd of November 2016.

To investigate this further the percentage availability for each node over the whole-time period is calculated and shown in Fig. 4 (left). The map shows that nodes 5, 11, and 17 are adjacent in the north-west corner of the site suggesting that these nodes were powered by the same supply and show the same power failures. They were also removed at the same time. Node 8 is situated at the southern end of the site. This node was installed but not powered for the initial 7 weeks resulting in the gap in the dataset. As these issues are known and a result of early removal or late power connection the availability percentages are recalculated and shown in the right-hand pane of Fig. 4.

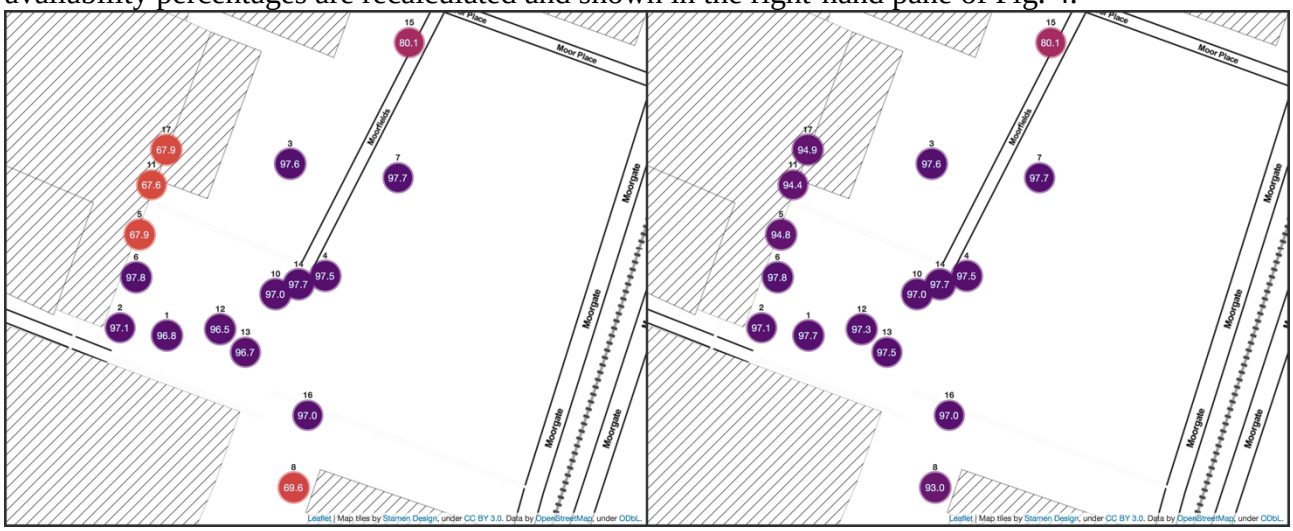

Figure 4: Data availability percentage map for all data (left pane) and corrected for late powering and early removal (right pane).

Heavy storms occurred in London on the 14th and 15th of July with several lightning strikes. This is the cause of the gap which occurs on all nodes. The erratic behaviour in node 15 is due to a lack of power during the daytime hours. The cause is uncertain but it could be the result of a change in site activities.

The use of a data availability matrix is a fast way to assess the performance of a network of sensors highlighting whether groups of nodes or singular nodes are missing periods of data.

3.2 Level Distributions

The $L_{A,eq}$ data for all nodes for the length of the campaign is shown in Fig. 5. Plotting this amount of data in the same visualisation masks individual trends due to the density of the data. In this case, this has been mitigated by reducing the size and opacity of the individual data points and highlighting the mean values. The daily and weekly cycle of peaks and troughs is visible, especially in the second half of the dataset. An extended gap can be seen that corresponds to the August bank holiday. The

background levels drop during the summer months (July-September). The loudest single events occur in the 2nd and 3rd week of July.

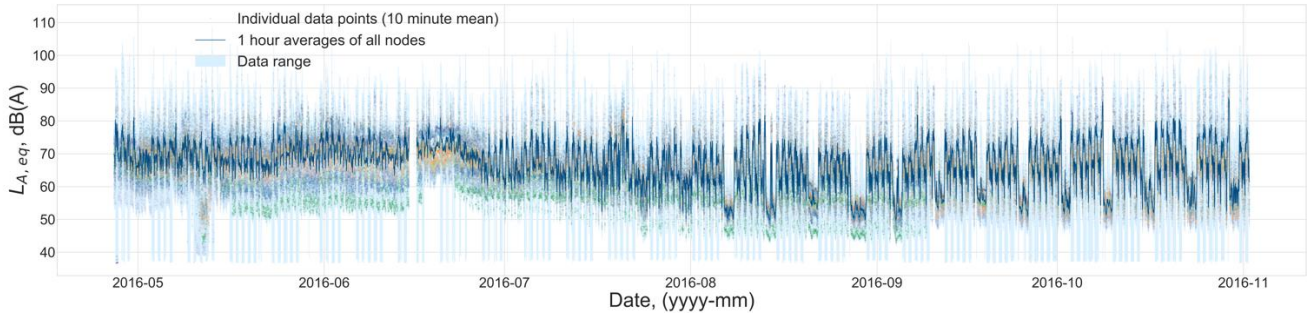

Figure 5: $L_{A,eq}$ data time series plotted as individual 10 minute averages (dots, coloured by node number), 1 hour averages of all nodes (dark blue solid line) and data range (pale blue area).

In July, the contractor for the test site took over responsibility for the neighbouring shaft site and began conducting activities during night time hours. To explore the impact of this, the distribution of the $L_{A,eq}$ data can be visualised using a split violin plot showing the kernel density estimation of the data, the median and the upper and lower quartiles. The data is split into two time ranges, 07:00 – 19:00 and 19:00 – 07:00.

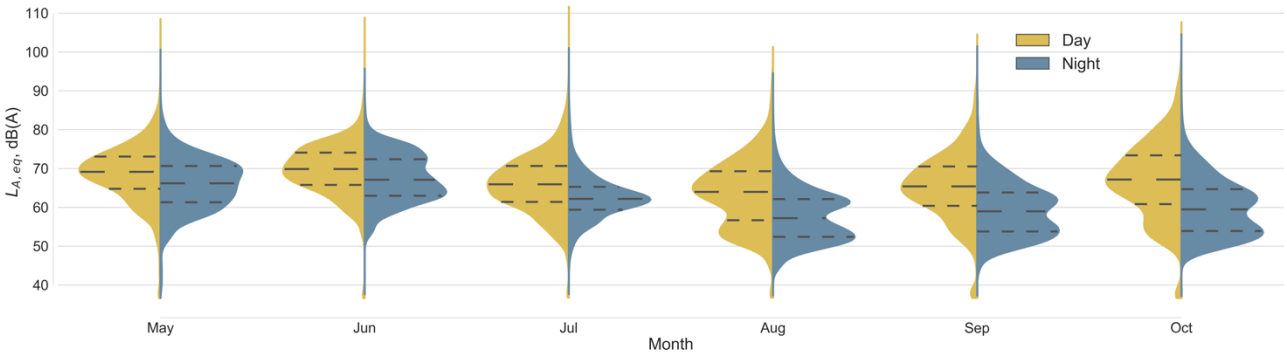

Figure 6: Day-night split violin plot of $L_{A,eq}$ for 6 months

Fig. 6 shows a day-night split violin plot for each full month of data generated. The expected impact of night time activities is an increase in the night time levels. However, this is not shown in Fig. 6. Instead, the night time levels reduce significantly from July onwards. The daytime levels follow the trend of Fig. 6 with a reduction during the summer months. In the night time distributions for August, September and October a double peak distribution is observed showing that there are two distinct noise distributions occurring. In this case, the louder peak is connected to the night time activity on the site and the lower peak is the noise levels which occur during several rest periods.

Plotting level distributions using a violin plot is an effective way to compare data from fixed periods especially when it is classified using binary categories such as day and night.

3.3 Events

Noise events are defined here as significant level increase above the ambient noise. A subset of the dataset can be taken by selecting the points of the time series in which one or more nodes measures a $L_{A,eq}$ above a threshold. For the following analysis, this threshold is set at 85 dB(A).

In Fig. 7 (left) the means and standard deviations of the levels found on all nodes are plotted as a scatter plot for all time points where one node has a value above 85 dB(A) and colour coded by the node with the highest level at that time. Most points have a mean between 75 and 85 dB(A) and a standard deviation between 3 and 8 dB. There are 3 distinct groupings outside this majority group to investigate further; one with mean levels near 90 dB(A) and standard deviations between 5.2 and

6.8 dB (upper-left), a second with high standard deviations but mean levels between 75 and 85 dB(A) (lower-right) and a third with high standard deviation and high levels (upper-right).

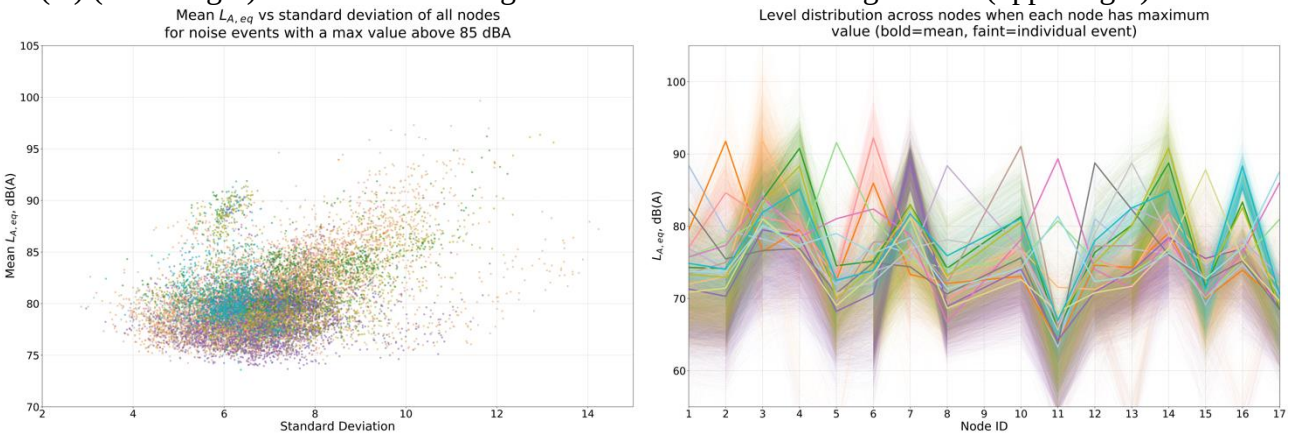

Figure 7: Mean $L_{A,eq}$ vs standard deviation (upper panel) and distribution across nodes for events where each node is peak (lower panel).

The right frame of Fig. 7 shows the mean (bold) and individual (faint) distributions when each node has the maximum value, highlighting groupings of nodes within the dataset. The density of faint lines shows the proportion of maximum values that occur at each node with nodes 3, 4 and 14 dominating in this case. Most neighbouring nodes have similar levels such as 4 and 14. However, despite its proximity to nodes 4 and 14, node 10 shows different behaviour. Nodes 4 and 6 appear to be opposites. Having identified several situations of interest the level distributions can be a mapped to show the spatial patterns present.

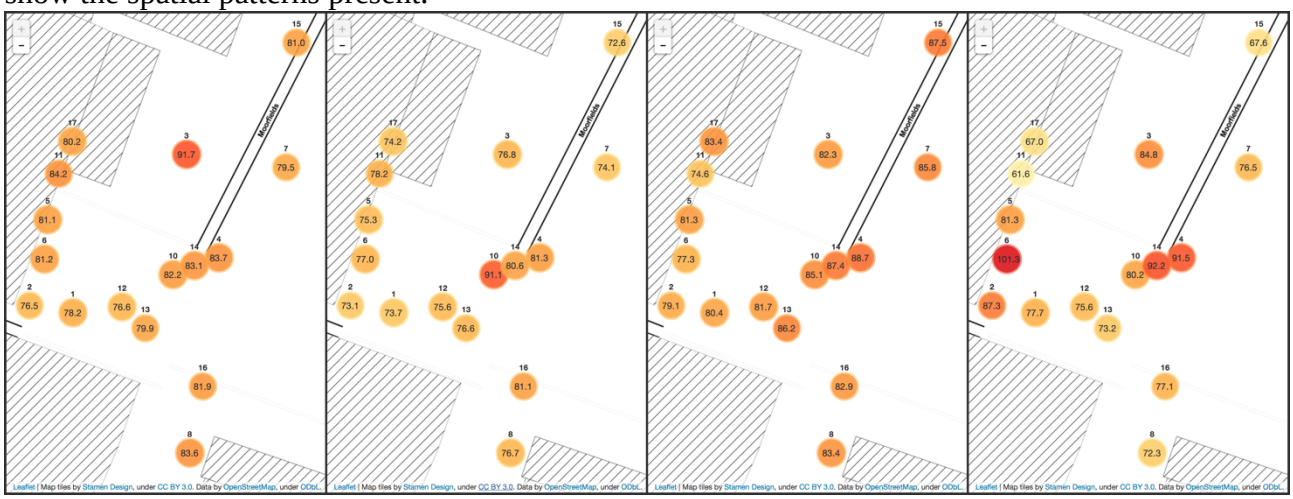

Figure 8: Mapped level distributions, mean distribution when node 3 is maximum (far-left), mean distribution when node 10 is maximum (mid-left), single event where standard deviation is low but peak is above 85 dB(A) (mid-right) and single event where nodes 4 and 6 have levels above 90 dB(A) (far-right).

Fig. 8 shows four cases – mean distribution when node 3 is the maximum, mean distribution when node 10 is the maximum, a single event where the standard deviation is low but the peak is above 85 dB(A) and a single event where both nodes 4 and 6 have levels above 90 dB(A). Examination of the patterns shown in Fig. 8 reveal that node 10 appears to measure noise in a unique situation compared to the other nodes, including its two nearest neighbours. It is one of the three nodes placed on the lower level overlooking the shaft. It is likely that events with peaks measured on this node are generated by activities on the shaft site with some degree of physical shielding from the other nodes and in close proximity. The far-right map in Fig. 8 shows two peaks with one on the shaft side and the other in the south west corner. This highlights that high standard deviation and high levels on nodes 4 and 6 can be used to identify events featuring two activities. This approach could be generalised to any pair of opposite nodes identified using the distributions shown in Fig. 7.

4. 4 days of 200 ms data

In this Section, the focus is on the level distributions and noise events found in datasets 2-5. The data availability for all nodes on all 4 days is above 99% and is not examined in detail here.

4.1 Level Distributions

Fig. 9 shows the time series for the 4 one day datasets. When compared to Fig. 5 more details are visible. On the 24th of June, the night time periods show a stratification suggesting a steady source of noise from outside the site, such as the road or the neighbouring constructions site. The 23rd of July features a single measurement above 100 dB(A) just after 06:00. It is noticeable that similar peaks are found on all 4 days suggesting that this event is the starting of a piece of equipment. On both the 18th and 20th of August activity during the middle of the day results in repeat readings near 80dB(A), the former appears less consistent whilst the latter features a steady noise on just one node.

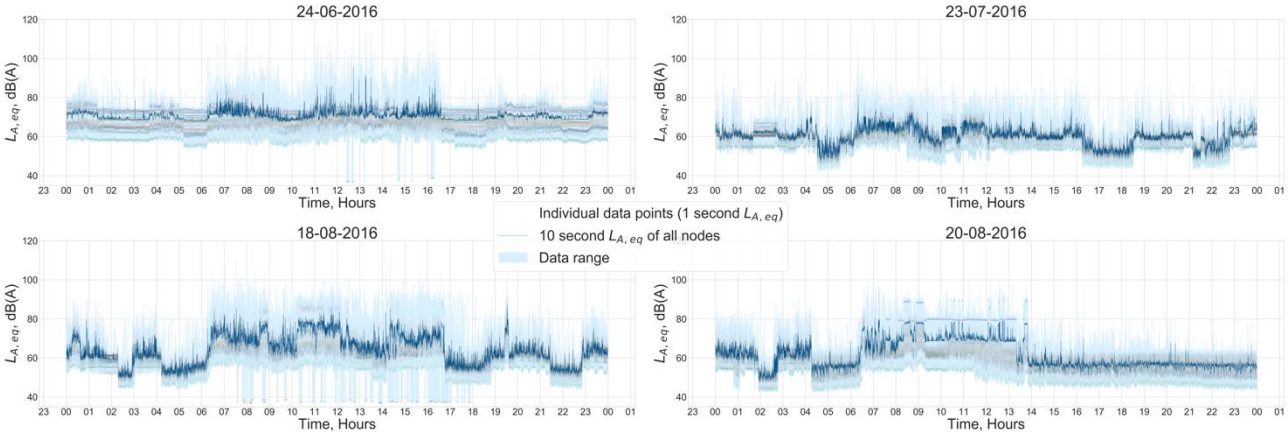

Figure 9: $L_{A,eq}$ time series plotted as individual 1 second averages (dots, coloured by node ID), 10 second averages of all nodes (dark blue solid line) and data range (pale blue area) for 4 separate days.

4.2 Events

The events highlighted in Section 4.1 are mapped in Fig. 10. In the first and last cases, a one hour mean is taken from the time ranges highlighted and in the other two cases a single event is shown.

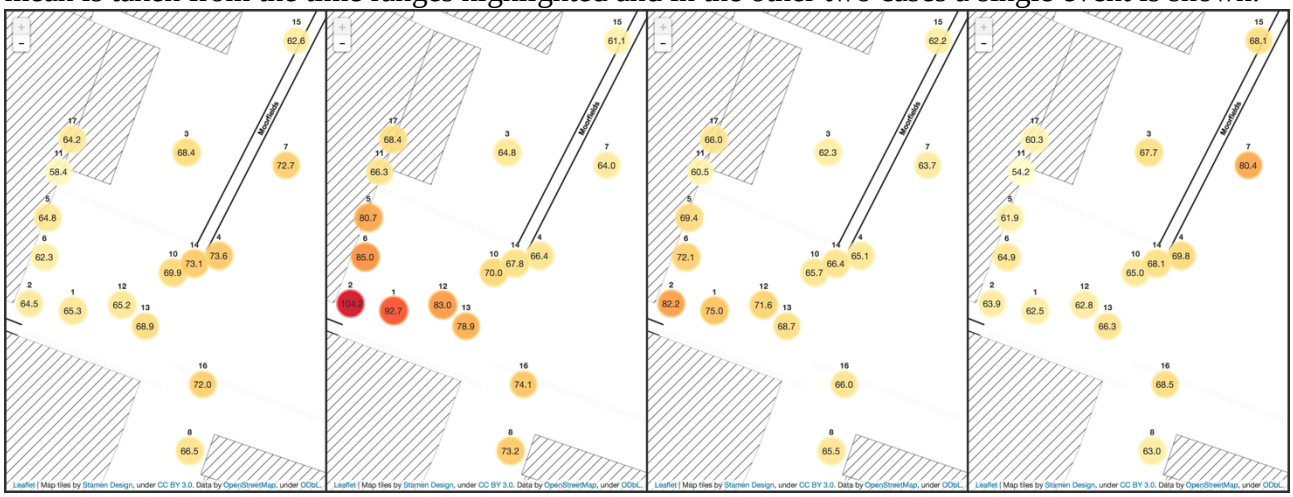

Figure 10: Mapped levels for events occurring on the 24/06/16, 23/07/16, 18/08/16 and 20/08/16

In all these cases the levels on most nodes are lower than those generally found in Fig. 8. This is due to shorter averaging periods or no averaging at all giving a more instantaneous impression. For the first case the higher readings are on the east side of the site and, due to the level drop of 8-10 dB, it is likely that the source is from within the shaft site rather than the main road for which a drop of less than 6 dB would be expected, assuming spherical spreading. The single event above 100 dB(A)

on the 23rd can be seen to affect the south west of the site. A similar pattern is found for the repeat events of the 18th although at a lower amplitude. On the 20th of August, the peak occurs at node 7 on the opposite side of the site and due to the distribution across the site, the source is likely to be local to node 7.

5. Conclusions

The construction site noise study undertaken in 2016 at Crossrail's Moorgate site generated a large and rich dataset containing many features. The study itself was highly focussed on separating the noise generated within the site from its surroundings, using spectral correlation methods.

In this paper, different analysis approaches have been applied to the dataset to reveal further features. These approaches have only dealt with the broadband level data. The data analysis and visualisation shown highlight both the challenges of working with a large noise dataset as well as some of the stories that are contained within. These stories include the gaps that occur due to electrical storms and installation issues, the changes in night time noise levels due to activity on the site and the changes of background noise during the summer months and the spatial distributions which occur due to activities in and around the site including how the topology of the site can contribute to peak in the measurements at particular locations.

The next steps which could be taken include adding the frequency data to the analysis to allow for the character of the noise to be assessed as well as the spatial variation and the application of machine learning tools, such as cluster analysis, for grouping patterns. By applying these methods, it is possible to record the types of activity that occur through analysis of the noise signatures. This can allow for more informative noise monitoring that can help to optimise projects, reducing noise impact and increasing efficiency.

Acknowledgements

The data used in this paper was generated by a project commissioned and funded as part of Crossrail's Innovate 18 program by Maggie Brown and was conducted by the scientists at the National Physical Laboratory. Installation and on-site support for the project was provided by Matthew Brinklow and John Castle on behalf of Laing O'Rourke. The authors would like to thank Crossrail for the permission to use the data.

REFERENCES

1. Control of Pollution Act 1974, *Her Majesty's Stationery Office*, London, 1974.
2. Sotirakopoulos, K., Barham, R., and Piper, B. Designing and evaluating the performance of a wireless sensor network for environmental noise monitoring applications, *EuroRegio 2016*, Porto, 2016.
3. Mydlarz, C., Salamon, J. and Bello, J.P. The implementation of low-cost urban acoustic monitoring devices, *Applied Acoustics*, **117** (B), 207-218, 2017
4. Van Renterghem, T., Thomas, P., Dauwe, S., Touhafi, A., Dhoedt, B., and Bottledooren, D. On the ability of consumer electronic microphones for environmental noise, *Journal of Environmental Monitoring*, **13** (3) 544,556, 2010
5. Alias, F., Socoró, J.C., Sevillano, X., and Nencini, L. Training an anomalous noise event detection algorithm for dynamic road traffic noise mapping: environmental noise recording campaign, *Proceedings of Tecni-Acústica 2015*, Valencia, October 2015
6. Alletto, F. and Kang, J. Soundscape approach integrating noise mapping techniques: a case study in Brighton, UK, *Noise Mapping*, **2**, 1-12, 2015
7. Stansfeld, S.A. Noise effects on health in the context of air pollution exposure, *International Journal of Environmental Research and Public Health*, **12** (10), 12375-12760, 2015
8. Sotirakopoulos, K., Piper, B., Barham, R. and Sheridan, S. Network-based noise monitoring of construction operations, Technical report, Prepared by the NPL for Crossrail Ltd, available at https://www.i3p.org.uk
9. Barham, R. The Secretary Of State For Business, Innovation & Skills Of Her Majesty's Britannic Government, *Microphone system and method*, World, WO 2013136063 A1, 19th September 2014