University of Strathclyde Science

Local Environmental Audit and Management System

Sampling review project

STATISTICAL
MATHEMATICAL
ADVICE
RESEARCH
TRAINING

Statistical analysis conducted by
Dr Alastair Rushworth
Statistical and Mathematical Advice, Research and Training (SMART)
Department of Mathematics and Statistics,
University of Strathclyde
Glasgow
G1 1XH

March 2017

Zero Waste Scotland commissioned this report for information purposes, and the content reflects the data and feedback collected, and the interpretation of the authors. It does not necessarily reflect the views of Zero Waste Scotland or Scottish Government.

LEAMS project analysis

Summary of the current and sampling scheme

Street indexing, sampling & zoning

Currently, roads and streets are the objects of sampling. Streets are assigned a unique index, with those > 500m in length subdivided into smaller sections. An Excel macro is used to select sample locations based on street index. Streets within each local authority (LA) are classified into one of 7 usage zones. A representative coverage of zones is achieved within local authorities, by sampling streets by zone in proportion to LA zone composition.

Site sampling frequency

Sites visited during 3 audits each year, 2 by the local authority and one by Keep Scotland Beautiful as independent verification. Each audit assess a randomly selected 5% of streets within each of 32 local authorities (13253 sites in 2014/15).

Data collected at sites visited

At each sampling site, information on several types of litter and environmental quality recorded using a categorical score, which is then converted to a numerical score. The information is converted into the binary classification acceptable or not acceptable for the site. LA performance is reported as proportion of sites that achieve acceptable status.

Summary of questions of interest

- Explore robustness and sensitivity of performance scores under current sampling. How sensitive are LA & national performance scores to a decrease in the number of samples?
- Explore possibility of obtaining representative ward level scores - would this require
- extra resources?
- How might LA and national scoring change if some zones were up-weighted in the sampling procedure - eg. areas of high traffic / footfall?
- How might sample locations be generated across LAs under proposed sampling schemes?
- In particular, zone classifications are likely to change in future and important to assess what change this could make to scoring consistency.
- How might repeat visits to tracts change the scoring, and how many repeat visits would needed to provide longitudinal insights?

Part 1. Considering reductions in current sampling

We'd like you to conduct a review of the way the existing dataset is used to create local and national scores. Would a smaller number of sample locations affect grading at local authority or national level, and, if so, by how much. We will make the street index and Excel macro used to draw the street samples, as well as historic grading data, available to inform this analysis.

When sites are surveyed, they are deemed either acceptable or not acceptable. Hence, we can assume that each site is a Binomial random variable and the total number of sites deemed acceptable within each council follows a Binomial distribution. This means that for a particular site, we expect it is as probable that the site will be acceptable as any other site in the same region.

When dealing with the performance estimate (percentage of acceptable sites) we can use the confidence interval equation for a binomial quantity to achieve a measure of accuracy. For example, the most commonly used formula for a binomial confidence interval is given by

$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$

where p is the proportion/percentage being estimated. The term being added to/subtracted from p defines the width or accuracy of this interval. This enables us to express how accurately a given sample size will recover the underlying ‘true’ performance. So for a given value of p, known to be the underlying true value, we can invert this expression to find the sample size required to measure the true performance with a given accuracy.

It is clear from the above example that the accuracy of an estimate obtained from a sample depends on both the underlying ‘true’ performance and the number of samples taken. As the number of samples increases, the quantity added to/subtracted from p will decrease, resulting in a narrow confidence interval and a more accurate measurement. Conversely, if the number of samples decreases, the quantity added to/subtracted from p will increase, resulting in a poorer performance estimate. It is also worth noting that when using the same number of samples, an underlying ‘true’ performance near to 50% will be more poorly estimated than performances near to 0% and 100%.

For each council being surveyed, the accuracy of the performance can be expressed as a confidence interval. The interval is a pair of values (a, b) such that with 95% probability, on repeated sampling, 95% of such intervals contain the true performance score.

If we assume that the values for the reporting period represent the true performance scores for the councils then it is possible to calculate the accuracy of the current sampling methodology (Table 1).

Council	Samples 15/16	Performance (%)	CI	Samples - 10%	CI	Proposal 1	CI	Proposal 2	CI
1	197	94.4	(90.55, 97)	177	(90.3, 97.11)	197	(90.55, 97)	197	(90.55, 97)
2	236	98.7	(96.65, 99.64)	212	(96.49, 99.67)	177	(96.2, 99.71)	118	(95.37, 99.79)
3	355	95.8	(93.29, 97.51)	320	(93.14, 97.58)	355	(93.29, 97.51)	355	(93.29, 97.51)
4	691	92.8	(90.65, 94.52)	622	(90.52, 94.6)	691	(90.65, 94.52)	691	(90.65, 94.52)
5	158	98.7	(96, 99.73)	142	(95.78, 99.76)	118	(95.37, 99.79)	79	(94.23, 99.86)
6	450	97.1	(95.25, 98.37)	405	(95.13, 98.42)	338	(94.9, 98.52)	225	(94.28, 98.74)
7	504	87.9	(84.84, 90.53)	454	(84.66, 90.66)	504	(84.84, 90.53)	504	(84.84, 90.53)
8	317	93.1	(89.87, 95.47)	285	(89.67, 95.58)	317	(89.87, 95.47)	317	(89.87, 95.47)
9	240	95.8	(92.73, 97.84)	216	(92.53, 97.92)	240	(92.73, 97.84)	240	(92.73, 97.84)
10	282	94.3	(91.17, 96.59)	254	(90.97, 96.69)	282	(91.17, 96.59)	282	(91.17, 96.59)
11	855	97.9	(96.76, 98.7)	770	(96.69, 98.73)	641	(96.55, 98.8)	428	(96.19, 98.95)
12	266	94	(90.64, 96.37)	239	(90.42, 96.48)	266	(90.64, 96.37)	266	(90.64, 96.37)
13	238	92.4	(88.56, 95.29)	214	(88.32, 95.42)	238	(88.56, 95.29)	238	(88.56, 95.29)
14	389	93.1	(90.21, 95.27)	350	(90.04, 95.37)	389	(90.21, 95.27)	389	(90.21, 95.27)
15	133	97	(93.01, 98.98)	120	(92.73, 99.04)	133	(93.01, 98.98)	133	(93.01, 98.98)
16	869	92.9	(91.02, 94.44)	782	(90.91, 94.52)	652	(90.7, 94.66)	434	(90.16, 95.01)
17	378	81	(76.76, 84.66)	340	(76.52, 84.85)	378	(76.76, 84.66)	378	(76.76, 84.66)
18	546	94.3	(92.14, 96.03)	491	(92.01, 96.11)	546	(92.14, 96.03)	546	(92.14, 96.03)
19	268	95.5	(92.54, 97.53)	241	(92.35, 97.61)	268	(92.54, 97.53)	268	(92.54, 97.53)
20	762	88.3	(85.9, 90.45)	686	(85.76, 90.56)	762	(85.9, 90.45)	762	(85.9, 90.45)
21	203	96.6	(93.35, 98.44)	183	(93.14, 98.51)	203	(93.35, 98.44)	203	(93.35, 98.44)
22	627	97.1	(95.6, 98.23)	564	(95.5, 98.28)	470	(95.32, 98.36)	314	(94.83, 98.57)
23	363	96.4	(94.13, 97.98)	327	(93.98, 98.04)	363	(94.13, 97.98)	363	(94.13, 97.98)
24	313	95.2	(92.42, 97.17)	282	(92.24, 97.26)	313	(92.42, 97.17)	313	(92.42, 97.17)
25	322	88.2	(84.34, 91.38)	290	(84.11, 91.53)	322	(84.34, 91.38)	322	(84.34, 91.38)
26	232	85.8	(80.85, 89.82)	209	(80.57, 90.01)	232	(80.85, 89.82)	232	(80.85, 89.82)
27	215	91.6	(87.37, 94.78)	194	(87.11, 94.92)	215	(87.37, 94.78)	215	(87.37, 94.78)
28	901	90.1	(88.04, 91.94)	811	(87.92, 92.03)	901	(88.04, 91.94)	901	(88.04, 91.94)
29	603	95.2	(93.26, 96.69)	543	(93.14, 96.76)	452	(92.92, 96.88)	302	(92.33, 97.19)
30	1005	97.9	(96.88, 98.66)	904	(96.82, 98.7)	754	(96.7, 98.76)	502	(96.37, 98.9)
31	1149	91.7	(90.03, 93.22)	1034	(89.93, 93.29)	1149	(90.03, 93.22)	1149	(90.03, 93.22)
Total	14067			12661		12866		11666	

Table 1: Accuracy of current sampling methods and proposed sampling methods.

Table 1 shows the number of samples used for each of the 31 council areas. It also shows the percentage of samples deemed acceptable along with a 95% confidence interval for that measure of acceptability. Essentially, this can be thought of as the values we would expect to see if we repeated the sampling again. We now assume that the performance scores given in column 3 represent the true underlying performance score.

The fifth column in the table shows the number of samples needed if we apply a 10% flat reduction in the sample size across all areas. Taking this new sample size together with the true underlying values given in column 3, we are able to re-estimate the confidence intervals for each of the regions. The resulting decrease reduces the total number of samples taken from 14,067 to 12,661. When comparing the two confidence interval estimates we see that for some areas, the reduction in sample size corresponds to a relatively small change to the width of the confidence interval. It is worth noting that the sample size change affects poor performing regions disproportionately.

An alternative to applying a flat reduction in sample size is to reduce the number of samples only for council areas that are performing well. The definition of performing well depends on the client's perspective but for this analysis, we have defined a performing well council to be one where, in 2015/2016, the width of the confidence interval is less than 3 percentage points (grey rows in Table 1).

Proposal 1 shows the results when there has been a 25% reduction in sample size for the performing well councils. This reduction means that the total number of samples has reduced from 14,067 in 2015/2016 to 12,579. Note that this new sample size is smaller than the one obtained from the 10% flat reduction. When we compare the confidence intervals for 2015/2016 with the ones calculated from the reduced sample size in proposal 1 we see that the values for the poor performing councils remain unchanged. Similarly, the width of the confidence intervals for the councils that are performing well changes only a relatively small amount. Meaning a relatively small decrease in accuracy.

Proposal 2 shows the results when there has been a 50% reduction in sample size for the performing well councils. This reduction means that the total number of samples has reduced from 14,067 in 2015/2016 to 11,091. When we compare the confidence intervals for 2015/2016 with the ones calculated from the reduced sample size in proposal 2 we again see that the width of the confidence intervals for the councils that are performing well changes only a relatively small amount. Meaning a relatively small decrease in accuracy.

Limitations

Our assumption that the total number of sites deemed acceptable within each council follows a Binomial distribution may not hold if zones within councils differ greatly in performance. This means that the calculated confidence intervals may be underestimated. That said, the principle of greater numbers of samples needed to estimate percentages nearer to 50% still holds.

Furthermore, our calculations rely heavily on the assumption that the acceptability scores for the reporting period 2015/2016 represent the true score for these areas. Since the scores for this reporting period are also estimates then it is possible that they differ from the true values. Given the consistency in estimates seen in the data provided during this project, we feel that this assumption does not present a major problem.

The definition of a council area that is performing poorly/well is very subjective. Here we have restricted the definition of poor to those council areas with confidence interval widths of length 3 or more percentage points. It is possible to extent this criteria to include more/less areas and this will have a large effect on the sample size. Therefore, the client needs to consider the amount of variability (given by the confidence interval width) it is willing to accept.

Part 2. Changing transect definition

It was not possible to quantify the effect of choosing different transect definitions using analysis without substantial trial data.

Part 3. Overweighting

We were asked to consider the possibility of overweighting for different zone types because certain zones should be more strongly represented in the overall score for a local authority.

In our opinion, there are two possible ways to achieve this weighting:

(1) Sample certain zones more intensively than others by considering the composition of the council area that currently determines the zone sample size and report the council score as the % of all sites that are acceptable. For example, if a council area is 80% urban and 20% rural, the sample size for that council area should be split 80% from the urban areas and 20% from the rural areas to ensure accurate representation.

(2) Sample zones in proportion to their prevalence in the council area (or based on a sample size calculation for each zone) calculated performance scores for each zone, and then weight and sum the zone scores to produce the council performance score. This would essentially assume a true underlying performance score for each of the zones that are to be estimated with a specific accuracy. Knowing the true score along with the desired level of accuracy would enable sample sizes to be determined using a similar argument to Part 1. The performance scores for each zone can be calculated and a weighted sum could then be taken to estimate the overall performance of the council area.

Option (1) is the most simple method of overweighting. That said, if key purpose of sampling is to estimate 'true' performance score, then to this end, the only reason to increase intensity would be to improve accuracy of this estimate, this suggest that option (2) would be more suitable.

The option (2) approach would ensure that individual zone performances are as accurate as desired, and the subjective weighting is applied after zone performances are obtain to produce the council performance. Following this argument, there is statistical justification to sampling more frequently in zone 1. Zone 1 generally performs more poorly, on average, than the others, and has

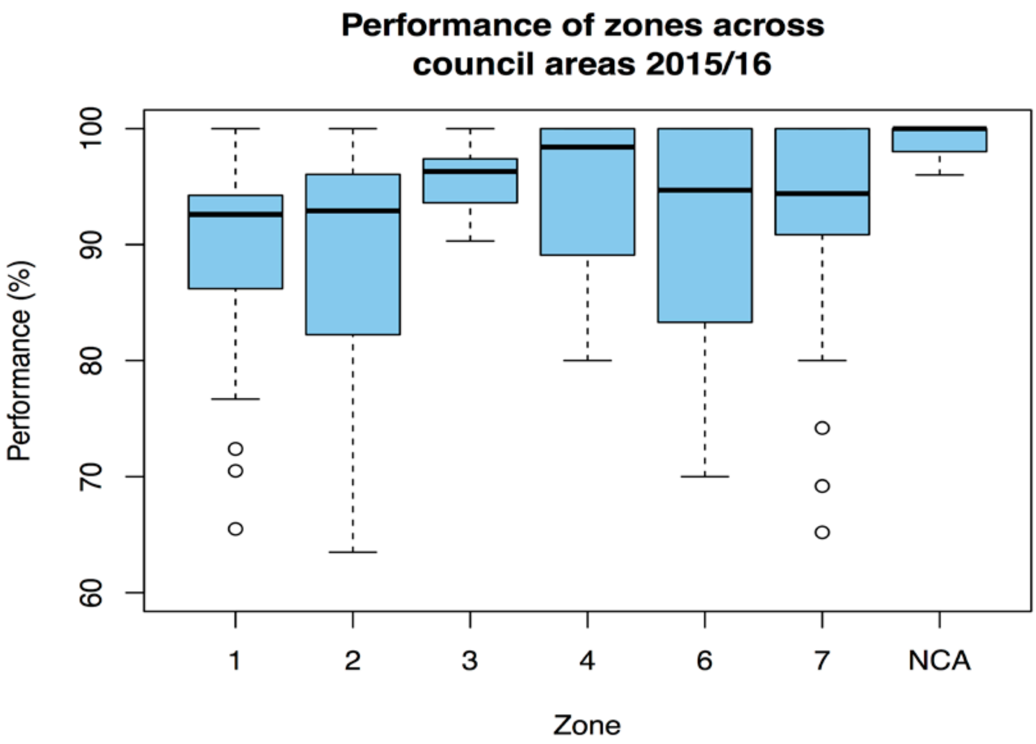


Figure 1: Boxplot showing the performance scores for the zones within council areas. Outlier values indicated by circles. The thick black horizontal line in the box represents the median

scores nearer to 50% than other less urban zones, and therefore requires a larger number of samples to estimate to the same level of accuracy as other zones (Figure 1).

Part 5. Sampling at ward level

The client was interested to know if it would be possible/viable to introduce a system to ensure a number of sample points covered each ward in a Local Authority area at every survey to allow local level results at ward level but still provide national results at zone level. Guidance was sought on the frequency of repeat sampling that provides a meaningful data set. The client was particularly concerned this might imply a very large number of sample points.

Data were unavailable for scoring at ward level, so it was difficult to quantify the likely number of samples needed. We are however, able to provide a very rough estimate.

There are 353 wards within Scotland (http://www.lgbc-scotland.gov.uk/faqs/councils.asp). For 2015/2016, the average performance by council was 93.63%. Without detailed data to inform the analysis, we must assume that each ward has an underlying performance of 93.63%.

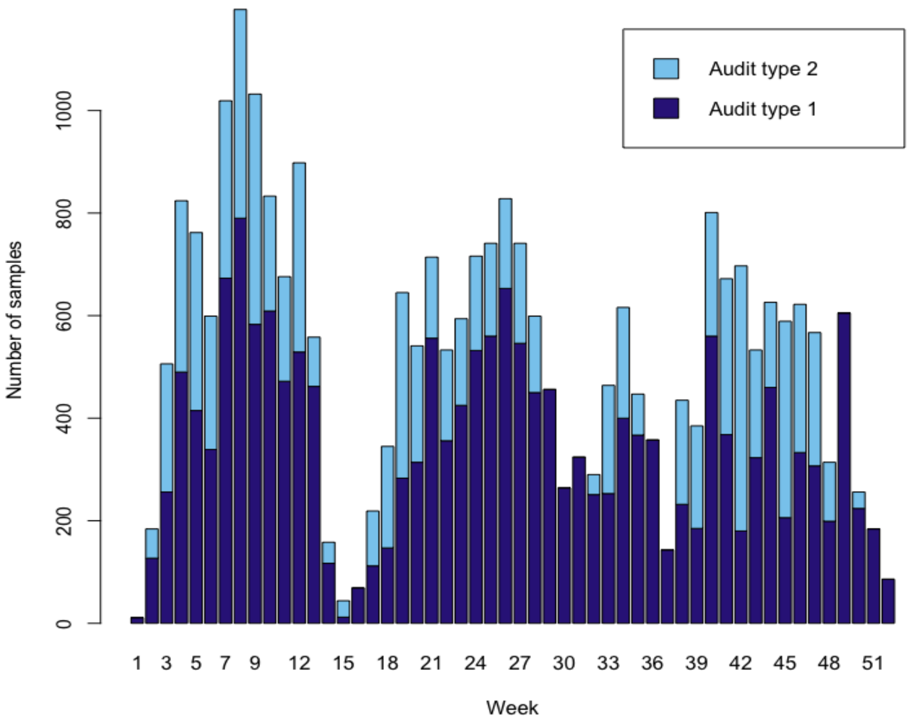
To estimate the ward performance within +/- 5% would require approximately 22,700 samples. Decreasing the accuracy of ward performance to within +/- 7.5% would require approximately 10,100 samples. Although both of these sample size estimates are crude they give a rough indication of the number of samples required to estimate ward performance.

The lack of available data at ward level meant that neither of these calculations take into consideration the possibility of obtaining representative coverage of zones at either council or ward level. It is likely that taking this into consideration would dramatically increase the number of samples required.

Part 6. Repeated sampling

The client was interested in knowing, from a statistical perspective, what implications might repeat sampling at some of the same locations have for the sampling regime and consistency of results? What balance of new and repeat transects might be desirable year to year?

There is no clear way that we can answer this statistically because it all depends of the purpose of repeat sampling. One goal of the repeat sampling process might be to assess whether a single site varies in quality, over time. This could be investigated using statistical modelling techniques, in particular Generalised Additive Modelling (GAM) which can capture complex patterns of independence over time with associated confidence measures. GAMs allow us to capture the seasonal and non-linear trends in data by modelling the dependence of the response on the predictors using smooth functions. A GAM splits the data range into separate sections, fits a simple curve to each section and joins them together at a series of points called knot points. Since the outcome of interest is in this instance is the site being acceptable or not acceptable we have a binary outcome. To model this kind of outcome, a logistic additive model could be used. In this kind of analysis you would could use zone, council and date of sample as predictor variables. Sampling is currently carried out across the whole year (figure below) so it is possible to use historical data to understand whether the time of year has an impact on the likelihood of an area scoring acceptable.



Audit type 2
Audit type 1
Number of samples
0
200
400
600
800
1000
1
3
5
7
9
12
15
18
21
24
27
30
33
36
39
42
45
48
51
Week