RMetS
Royal Meteorological Society

Geoscience Data Journal
Open Access

Missing data in spatiotemporal datasets: the UK rainfall chemistry network

J. N. Cape*, R. I. Smith and D. Leaver

Centre for Ecology & Hydrology, Edinburgh, UK

**Correspondence: J. N. Cape, CEH Edinburgh, Bush Estate, Penicuik, EH26 0QB, UK, E-mail: jnc@ceh.ac.uk*

This article was funded by the UK Department of the Environment, Food and Rural Affairs and Devolved Administrations.

Rainfall chemistry networks inevitably report some missing data, caused by contamination or loss of samples. However, there are no universally accepted rules about how such data, particularly from samples contaminated in the field, are identified and reported, leading to uncertainties in data usage by third parties, and possible incorrect inferences based on the reported data. This paper describes how the UK rainfall chemistry network data have been analysed for contamination, and how missing values can be estimated based on cross-correlations in time and space, using data from 20 sites over 26 years. The final flagged dataset is available through the CEH Environmental Information Data Centre (EIDC). Erroneous data values are identified through consideration of ion balance (internal consistency), and evidence of contamination by birds or windblown dust based on the reported chemical analysis. Overall data capture with the erroneous data excluded and no replacement of missing data was 86%, but with much smaller data capture at some sites in some years, to <30% in some cases. The use of estimated data to replace missing values resulted in an increase in overall data capture to 96%, with only one site having data capture <70% in an individual year, and all sites achieving a data capture of 88% or more over the full period. The implications of using the reported 'official' annual data, as opposed to the dataset with missing values replaced by estimates, are illustrated by consideration of the temporal trend in nitrate at one site, which shows twice the value in the 'official' reported annual data compared with the 'estimated' data, part of a consistent pattern across all sites. Use of the uncorrected 'raw' sample data leads to large errors.

Geosci. Data J. **2***: 25–30 (2015), doi: 10.1002/gdj3.24*

Received: 18 August 2014, revised: 9 February 2015, accepted: 18 February 2015

Key words: missing values, temporal trends, contaminated data

Dataset

Identifier: doi: 10.5285/ada39609-ddec-4cbe-85c2-4fdd6bd774d7; CEH:EIDC:1398956265930
Creator: Cape, J. N.
Title: Cleaned UK Rainfall Chemistry Data (1986–2011)
Publisher: EIDC (gateway.ceh.ac.uk)
Publication Year: 2014

Introduction

The phenomenon of 'acid rain', a term first coined at the end of the nineteenth century (Smith, 1872), sparked an interest in the chemical composition of rainfall, and the consequent deposition of acidifying pollution in areas remote from pollutant sources (Eaton *et al.*, 1980). In the four decades since the potential for environmental damage was recognized, many networks for measuring rain chemistry have been developed, e.g. the European Monitoring and Evaluation Programme (Tørseth *et al.*, 2012). The large-scale reduction in emissions of pollutants such as sulphur dioxide over the past 20 years, particularly in Europe, has been reflected to some extent in the network measurements of precipitation chemistry, but the reductions in deposition have not necessarily matched the reductions in emission in either space or time (Fowler *et al.*, 2007). At a global scale, in regions where emissions have been growing, new networks are being developed to monitor deposition and assess the potential for adverse effects, but there are still many regions with inadequate data to describe the chemical composition of precipitation (Vet *et al.*, 2014).

Rainfall chemistry sampling networks inevitably have some missing data values, because of equipment malfunction, accident, vandalism, or contamination. Efforts to improve data quality have led to several designs of both 'bulk' samplers (continuously open) and 'wet-only' samplers that open only when rain is detected (Galloway and Likens, 1976, 1978; Erisman *et al.*, 2003). Where 'bulk' deposition is measured with continuously open sampling funnels, the funnels are used as bird perches, despite guards and deterrents or the provision of alternative perches nearby, and the resultant fouling destroys the chemical integrity of the sample. Other sources of contamination are wind-blown dust, particularly in dry weather with disturbance such as harrowing of fields or quarrying, or resuspension from unmetalled roads. Total loss of a sample is rare, but contamination occurs not infrequently, particularly for bulk samplers as used in the UK National Rainfall Chemistry Network (PrecipNet: http://uk-air.defra.gov.uk/networks/network-info?view=precipnet).

Identification of contamination by birds is relatively straightforward; such samples are characterized by high concentrations of phosphate, ammonium, and/or potassium. Dust contamination is less easily identified, but would (e.g.) be seen in very high concentrations of calcium in the absence of sea-salts. Quality control measures to eliminate contaminated samples from the subsequent analysis of the data are routinely applied to rainfall network data. Different networks and researchers use slightly different approaches, but the basic principles are similar:

1. Is there evidence from field notes, or visible evidence, of gross contamination? This type of contamination may lead to rejection of a sample even before chemical analysis.
2. Is the chemical analysis of the water sample internally consistent, e.g. is there a charge balance between anions and cations? If not, samples may be sent for re-analysis if a problem in the laboratory is suspected. Clues to the source of a discrepancy may be obtained from comparison of the theoretical and measured conductance, or from the ratios of sea-salts (chloride:sodium).
3. Is there evidence of contamination from birds or dust on the basis of the chemical analysis?

Detailed guidelines for recording possible contamination in the field, and for quality assurance of laboratory data, have been published by the World Meteorological Organisation Global Atmosphere Watch (WMO/GAW: www.wmo.int/gaw/, Allan, 2004). However, these guidelines do not include methods for identifying non-visible contamination, for example from dust or bird contamination, in samples which pass the laboratory quality assurance criteria. These guidelines refer to 'wet-only' samplers, which are less likely to be contaminated than the continuously open 'bulk' samplers used in the UK network.

This type of quality control procedure leads to the elimination of data from such samples in the statistical analysis of the dataset, and the generation of 'missing' data. The importance of missing values in long-term datasets and the possible bias introduced by missing data, has been addressed recently for network sampling of airborne polycyclic aromatic hydrocarbons (PAH) (Brown, 2013). The practice in PrecipNet for weekly/two-weekly precipitation samples has been to assign to missing data the average for the dataset for the year, i.e. the annual average concentration is based on the valid data samples. For example, in a year with 26 sampling occasions, for which data from two are 'missing', the average (rainfall weighted) concentrations of the remaining 24 are used to replace the missing data in calculating the annual average concentration and deposition. The annual PrecipNet report (e.g. http://uk-air.defra.gov.uk/reports/cat13/1105130856_UKEAP_report_2010_Final.pdf) provides summaries of corrected data for each site, i.e. replacing contaminated sample data with the annual average of the remaining data. However, researchers wishing to examine the data in more detail must download the raw data from the UK-Air database (http://uk-air.defra.gov.uk/data/), and then apply their own methods for dealing with contamination and missing data. Unfortunately, the raw data are not flagged as potentially contaminated, so users may inadvertently calculate erroneous annual data that are at variance with the published data in the annual reports.

This study investigated whether 'missing' data lead to bias in the calculation of annual average concentrations and deposition, and in the identification of time-trends. Methods are proposed for reducing such uncertainties by estimating values for missing data on the basis of the whole available dataset, using both within-site and between-site correlations to model the whole data structure.

1. Data Production Methods

The raw data for this study were obtained from the UK-Air database, for each of the PrecipNet sites currently operating that have at least 23 years of data available. For most sites, sampling started in 1986, and data were assessed up to the end of 2011, a period of 26 years. Site details may be obtained from the UK-Air website (http://uk-air.defra.gov.uk/interactive-map?network=precipnet). Note that most of these 'bulk' data have not been reported to international databases, such as the EMEP database (http://ebas.nilu.no). For those sites where 'bulk' data have been reported, most appear to have been flagged as 'contaminated' where contamination is evident in the data. No data flags to denote contamination or other information are currently used by the UK-Air database.

1.1. Elimination of low-quality data

A consistent set of rules, based on observation of the complete dataset, was used to eliminate low-quality data from the UK-Air raw data:

1.1.1. Ion balance

The ion balance was calculated as (Σcations − Σanions)/(Σcations + Σanions). Some imbalance may occur because not all ions are analysed, e.g. bicarbonate or weak organic acids, so that a small positive balance is usually observed. Data were regarded as acceptable where the ion balance lay in the range −10% to +20%, except where the overall ion concentration (Σcations + Σanions) was <200 μeq l^{-1}, where the uncertainties in individual ion analyses at low concentrations could lead to a greater uncertainty in calculating the ion balance; in such cases the range was extended from −10% to +30%. These criteria differ slightly from those used by WMO/GAW (Allan, 2004), in that the threshold in overall ion concentration for accepting ion balances up to +30% is more lax (200 μeq l^{-1} rather than 100 μeq l^{-1} in WMO/GAW). However, the criterion for rejecting samples with an excess of anions over cations is more strict; an ion balance of −10% applies over all samples in this study, while ion balances up to -20% are regarded as valid but flagged as low quality by WMO/GAW.

1.1.2. Evidence of contamination

Samples with measured phosphate concentrations above 10 μeq l^{-1} were eliminated. All samples with ammonium concentrations above 100 μeq l^{-1} were examined, and those with potassium concentrations above 8 μeq l^{-1} were eliminated. Finally, all samples with calcium concentrations greater than 50 μeq l^{-1} which had higher calcium concentrations than sodium concentrations (in μeq l^{-1}) were excluded as being contaminated with wind-blown dust. This last condition is rather conservative, i.e. some dust-contaminated samples may remain in the dataset; these usually were samples with very low sample volumes, so contributed little to the annual deposition.

1.1.3. Missing data

Small sample volumes were not chemically analysed, but contribute little to the overall deposition of ions during a year. A small number of larger-volume samples had missing data, usually for either anions or cations, or occasionally where an ion (usually calcium) was below the limit of detection and recorded as 'missing' rather than 'zero'. All the pH data for 2000 are recorded in UK-Air as 'missing' because of doubts about the validity of the analysis in that year.

1.2. Estimation of individual missing data from samples

For samples in category 1.1.3. (above) where only one or two values were missing from an otherwise complete analysis, the missing data were estimated, for each site separately, by using the correlations between one ion and all the others at that site, over the 26 years, to estimate the missing values. This was achieved by using the GenStat (GenStat Release 13.1, PC/Windows XP, VSN International Ltd.) procedure MULTMISSING[1] on the log-transformed deposition data (i.e. logarithm of the product of concentration and rainfall amount). This method was used to avoid the highly skewed nature of the concentration data, with very high concentrations at low rainfall volumes. The resultant estimated values were used to replace missing data in the quality-controlled dataset. With the exception of the pH data for 2000, most values estimated by this technique were of small concentrations close to (or below) the reported analytical limit of detection. The 'reconstructed' samples were subjected to the same quality control procedures as full samples (i.e. ion balance, contamination), and rejected if the criteria were not met.

1.3. Estimation of missing data because of contamination or anomalous ion balances

The data for all 20 sites and 26 years, for each ion separately, were used to estimate missing values from excluded samples, by using the Genstat Procedure MULTMISSING, as described above. This method utilizes the correlations in time and space between sites to estimate missing data. Different sampling dates have been used at the different sites, so temporal matching was achieved by allocating the mid-point of the sampling period (weekly or 2-weekly, depending on year) to the nearest week number (i.e. no. of weeks from 1 January 1986). On some occasions there were too many missing data for an estimate to be made, and these samples remained as 'missing'.

The chemical composition of the 'missing' samples was therefore built up independently for each ion, and then the final composition of a 'missing' sample was assessed for compliance with the data quality criteria. In most cases, an acceptable ion balance was achieved, providing some confidence in the method used, but some samples also failed the assessment for contamination.

Finally, annual deposition estimates were made: (1) based on only the 'clean' samples with individual values estimated as in 1.2. above, i.e. using the measured annual rainfall to scale the deposition based on the valid data in a given year, and (2) based on the 'estimated' data, i.e. using the measured annual rainfall to scale the deposition based on the deposition including valid estimated values for missing whole records. These latter datasets also contained several 'missing' data for which estimates could not be obtained, but a much smaller fraction than for the 'clean' datasets.

[1] MULTMISSING uses an iterative regression technique to estimate missing values in a multivariate dataset, based on multiple linear regressions among all the variates; the default 10 iterations were used here. Further details can be found at http://www.vsni.co.uk/software/genstat/htmlhelp/server/MULTMISS.htm.

1.4. Description of the final dataset

Table 1 shows the proportion of data in the 'clean' (i.e. non-contaminated) datasets and the 'estimated' datasets, expressed in terms of the volume of precipitation represented by the data. Although the 'clean' datasets overall represent a large proportion of the precipitation volume (on average over 86%), individual years may be poorly represented. There is a marked improvement in data capture for the 'estimated' datasets, to an overall average of 95.7%, with only one site (Loch Dee) falling below 90% data capture. There is also evidence for better rates of data capture over time, particularly from 2003 after a full review of the network. The large difference in some years (e.g. 1988) between the 'official' and 'clean' datasets is caused by the identification and removal of samples contaminated by wind-blown dust, which is not considered in presenting the 'official' data in the published reports.

This difference in data capture is also reflected in the annual average deposition and concentration data for each site. For example, across all sites, average wet N deposition (g m^{-2} year^{-1}) shows a consistent effect of using the 'estimated' dataset; at low deposition rates the estimated data show higher values than the 'official' data, because where the raw data show a zero or missing value for ammonium concentrations (below detection limit), the estimated data give small positive values close to the detection limit. At high deposition rates, the estimated data show lower values than the 'official' data because more contaminated samples have been identified and excluded from the analysis. Uncritical use of 'raw' data direct from UK-Air gives very misleading results because the raw data include contaminated samples.

For trends (linear slopes of annual deposition across years), across all sites, the sea-salts show no difference between 'official' and 'estimated' data – this is not surprising as few samples are either contaminated or below the detection limit. For calcium, the trends from 'official' data are about half those seen in the 'estimated' data – reflecting the removal of 'dust contaminated' data from the 'estimated' dataset to give a less noisy signal. For ammonium, nitrate and non-sea sulphate, the 'official' data show a larger trend than the 'estimated' data across all sites (with some variability between sites in both absolute trends and comparison with 'estimated' data). For non-sea sulphate this is small, about 12% difference in trends overall; for nitrate and ammonium the difference is bigger, closer to 30%. An example is shown in Figure 1, where the annual average nitrate concentrations from this study are compared against the 'official' published data in the PrecipNet annual reports, and with the 'raw' and 'clean' data. In this case there is a significant difference between the datasets, leading to a factor of two difference in the simple linear trend over time at this site between the 'official' and 'estimated' data.

This is not an isolated example; Figure 2 shows the trend slope in annual average nitrate concentrations (μeq l^{-1} year^{-1}) for each of the sites based on

Table 1. Data capture averaged over 26 years for each site, expressed as the volume of valid samples relative to total volume: % 'clean' includes valid samples and those with individual data estimated, % 'estd' includes all samples with invalid data replaced (where possible). Min % denotes the minimum data capture in the 26-year period.

Site	% 'clean'	Min %	% 'estd'	Min %
Allt a'Mharcaidh	80.5	40.2	95.5	70.5
Bannisdale	93.8	52.3	99.5	97.4
Barcombe Mills	81.6	26.0	96.1	80.3
Bottesford	86.3	58.6	93.7	70.6
Eskdalemuir	91.6	73.1	98.8	94.3
Flatford Mill	82.6	54.2	92.8	72.4
Goonhilly	87.6	51.6	96.6	86.3
High Muffles	93.1	78.7	98.1	90.8
Hillsborough	81.8	58.7	95.2	82.1
Loch Dee	74.8	23.9	88.0	40.8
Lough Navar	93.2	70.4	97.9	87.1
Preston Montford	86.3	66.3	96.4	79.3
Pumlumon	88.6	66.6	97.6	84.1
Stoke Ferry	80.8	59.7	95.3	80.6
Strathvaich	98.6	88.6	99.0	88.6
Thorganby	74.2	26.5	93.5	83.9
Tycanol Wood	89.9	69.6	97.5	76.1
Wardlow Hay Cop	79.0	48.9	90.9	80.7
Whiteadder	91.7	71.5	94.2	75.9
Yarner Wood	92.7	75.7	97.7	79.2
Average	**86.4**		**95.7**	

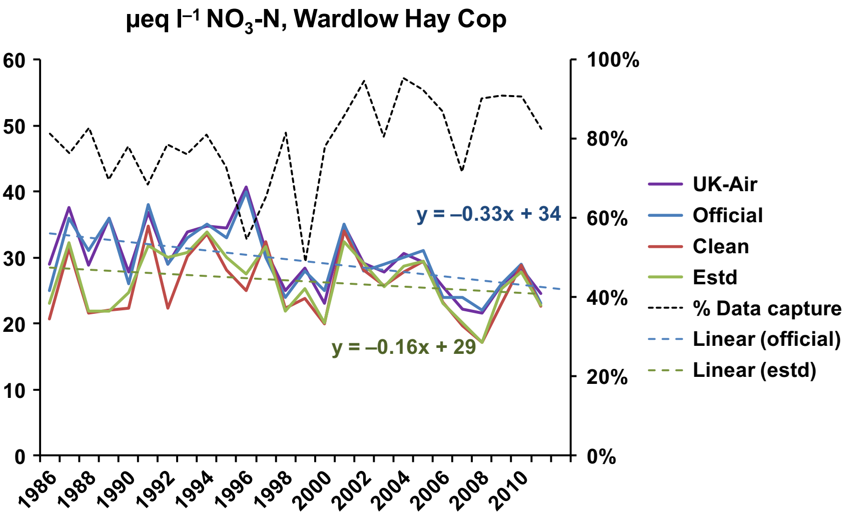

Figure 1. Comparison of annual average nitrate concentration in precipitation at Wardlow Hay Cop. 'UK-Air' are values calculated from the raw online dataset (http://uk-air.defra.gov.uk/data/), 'official' is the published value in PrecipNet reports (e.g. http://uk-air.defra.gov.uk/reports/cat13/1105130856_UKEAP_report_2010_Final.pdf), 'clean' is the estimate based on valid sample composition (as defined in text), and 'estd' is the estimate with missing values replaced (as described in text). The 'clean' data represent 79% of total volume over the 26 years, and 'estd' data represent 91% of the sampled volume.

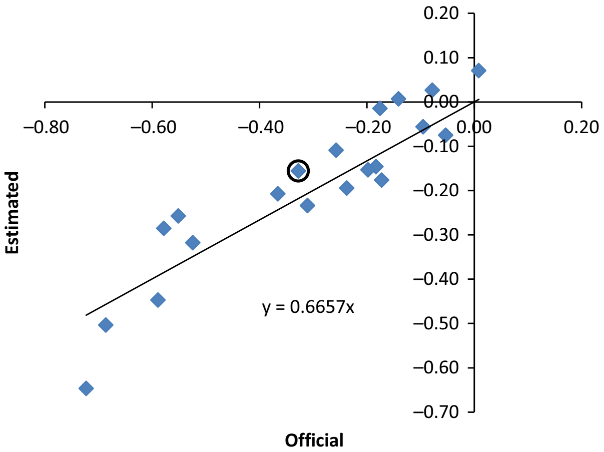

Figure 2. Plot for all sites of the long-term linear trend in annual nitrate (NO_3) concentrations (μeq l^{-1} year^{-1}) derived from the estimated data in the dataset described here and the 'official' statistics (e.g. http://uk-air.defra.gov.uk/reports/cat13/1105130856_UKEAP_report_2010_Final.pdf).

'official' or 'estimated' data. The data point for Wardlow Hay Cop (the example in Figure 1) is circled; on average the long-term trends derived from the estimated data are only 67% of the values derived from the 'official' annual data.

2. Dataset location and format

The final dataset, flagged as appropriate (valid, individual values estimated, invalid and replaced by estimated values, invalid and excluded) for all sites and years is available through the CEH Gateway (http://gateway.ceh.ac.uk) as an 'added value' product. Data are stored in comma separated variable format by site, with one row of each data table corresponding to a single sample date. The data are accessible through DOI 10.5285/ada39609-ddec-4cbe-85c2-4fdd6bd774d7.

3. Dataset use and reuse

Use of the UK-Air data 'as is' by researchers and consultants downloading data from UK-Air without an appreciation of the nature of the dataset (i.e. including contaminated samples which are not identified from standard laboratory quality assurance assessments) will lead to erroneous estimates of wet deposition. If annual-average data from UK annual reports are used, which have had most of the bird-contaminated data values removed, the error is reduced, but not entirely removed, because the annual averages are calculated based solely on the chemical composition of uncontaminated samples. The published annual average data take no account of sources of contamination other than birds (as evidenced by the presence of phosphate in the sample), such as wind-blown dust.

The provision of a dataset with data flagged as 'contaminated' or otherwise 'missing' will be of greatest benefit to the research community only if a clear link is provided from the UK-Air database.

Examples are given above of the scale of uncertainty introduced by comparing the 'official' data (based on the reported annual averages in PrecipNet reports) with data where 'missing' values are replaced using the best estimate based on cross-correlations in space and time. Even the relatively small changes in annual average data can, however, for some ions at some sites lead to large differences in predicted temporal trends. Rigorous application of criteria for identifying the contamination of 'bulk' samplers leads to a greater number of 'missing' data, and emphasizes the need to provide estimates in order to avoid bias in calculating annual deposition.

Acknowledgements

The research presented in this paper was funded by the UK Department of the Environment, Food and Rural Affairs and Devolved Administrations (Defra). The authors are grateful to Keith Vincent (Ricardo-AEA) for extracting the multi-site, multi-year datasets from UK-Air.

References

Allan MA (ed.) 2004. Manual for the GAW Precipitation Chemistry Program. WMO/GAW Report No. 160 ftp://ftp.wmo.int/Documents/PublicWeb/arep/gaw/gaw160.pdf (accessed 5 February 2015)

Brown RJC. 2013. Data loss from time series of pollutants in ambient air exhibiting seasonality: consequences and strategies for data prediction. *Environ Sci Processes Impacts* **15**: 545–553, doi:10.1039/c3em30918e.

Cape JN. 2014. Cleaned UK Rainfall Chemistry Data (1986–2011). EIDC (gateway.ceh.ac.uk). doi: 10.5285/ada39609-ddec-4cbe-85c2-4fdd6bd774d7; CEH:EIDC: 1398956265930.

Eaton JS, Likens GE, Bormann FH. 1980. Wet and dry deposition of sulfur at Hubbard Brook. In *Effects of Acid Precipitation on Terrestrial Ecosystems*, Hutchinson TC, Havas M (eds). Plenum Press: New York; 69–75.

Erisman JW, Mols H, Fonteijn P, Geusebroek M, Draaijers G, Bleeker A, van der Veen D. 2003. Field intercomparison of precipitation measurements performed within the framework of the Pan European Intensive Monitoring Program of EU/ICP forest. *Environmental Pollution* **125**: 139–155, doi:10.1016/S0269-7491(03)00082-4.

Fowler D, Smith R, Muller J, Cape N, Sutton M, Erisman JW, Fagerli H. 2007. Long term trends in sulphur and nitrogen deposition in Europe and the cause of non-linearities. *Water Air Soil Pollution: Focus* **7**: 41–47, doi:10.1007/s11267-006-9102-x.

Galloway J, Likens G. 1976. Calibration of collection procedures for the determination of precipitation chemistry. *Water, Air, and Soil Pollution* **6**: 241–258, doi:10.1007/bf00182868.

Galloway JN, Likens GE. 1978. The collection of precipitation for chemical analysis. *Tellus* **30**: 71–82.

Smith RA. 1872. *Air and Rain. The Beginnings of a Chemical Climatology*. Longmans, Green & Co: London.

Tørseth K, Aas W, Breivik K, Fjæraa AM, Fiebig M, Hjellbrekke AG, Lund Myhre C, Solberg S, Yttri KE. 2012. Introduction to the European Monitoring and Evaluation Programme (EMEP) and observed atmospheric composition change during 1972–2009. *Atmospheric Chemistry and Physics* **12**: 5447–5481. doi:10.5194/acp-12-5447-2012.

Vet R, Artz RS, Carou S, Shaw M, Ro C-U, Aas W, Baker A, van Bowersox C, Dentener F, Galy-Lacaux C, Hou A, Pienaar JJ, Gillett R, Forti MC, Gromov S, Hara H, Khodzher T, Makowald NM, Nickovic S, Rao PSP, Reid NW. 2014. A global assessment of precipitation chemistry and deposition of sulfur, nitrogen, sea salt, base cations, organic acids, acidity and pH, and phosphorus. *Atmospheric Environment* **93**: 3–100, doi:10.1016/j.atmosenv.2013.10.060.