Statistics New Zealand
TATAURANGA AOTEAROA

Assessing data quality in official environmental statistics

Adam Tipper, Senior Analyst, Statistics New Zealand

Presentation to the UN Expert Group on Environmental Statistics, April 22 2016

Rationale for assessing environmental data quality

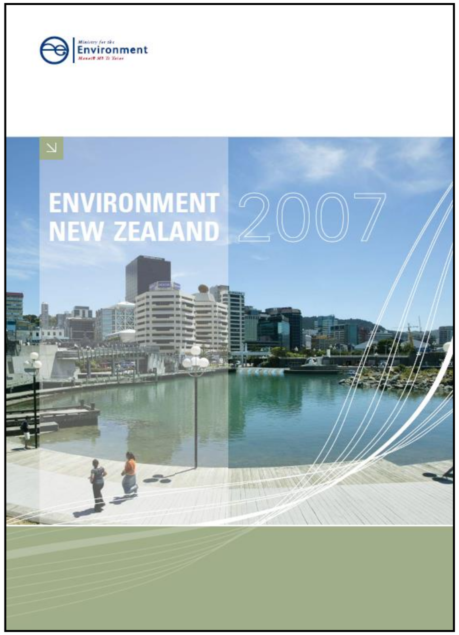
- Enable maximum efficient and proper use to be obtained from the data.
- Range of customers need environmental data, from national-level policy decision makers, to local government, researchers, businesses, the general public, for a range of reasons.

Overview

- Environmental Reporting in New Zealand.
- The data quality challenge.
- Testing the Data Quality Framework against environmental data.
  - Process
  - Key issues
- Summary and discussion.

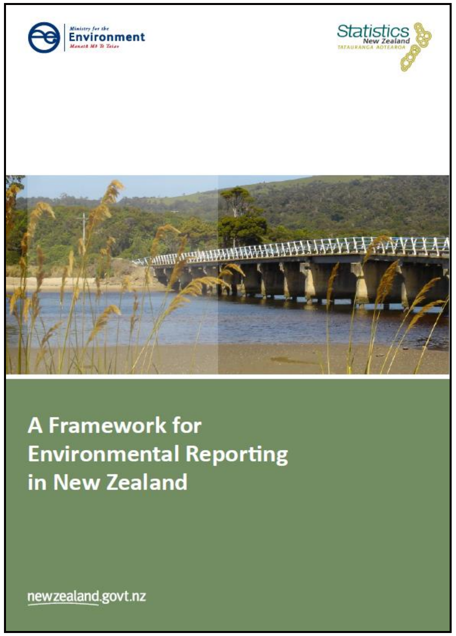
Environmental Reporting in New Zealand

2007 **2014** **2015**



Environment

ENVIRONMENT NEW ZEALAND 2007

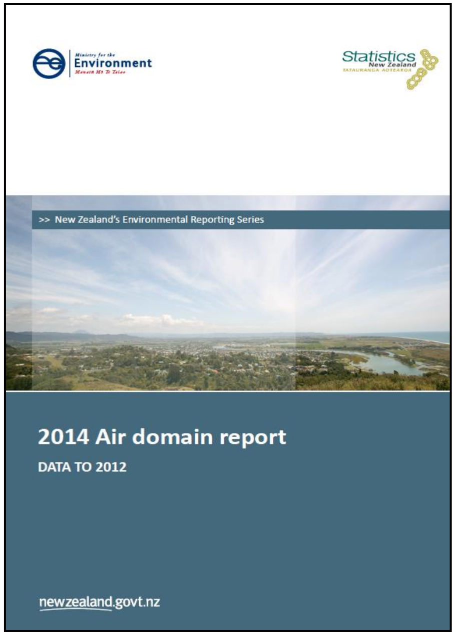


Environment

Statistics New Zealand

A Framework for Environmental Reporting in New Zealand

newzealand.govt.nz



Environment

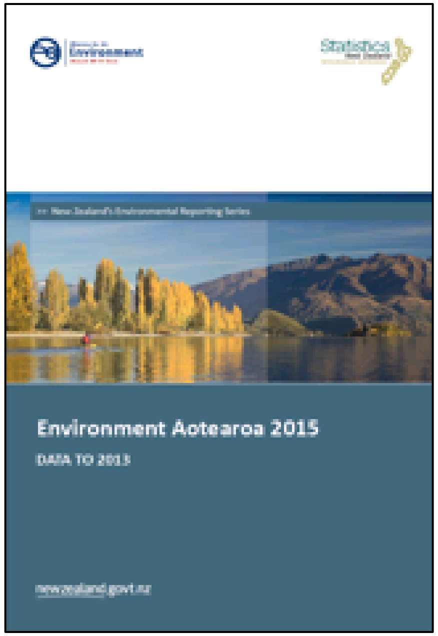
Statistics New Zealand

>> New Zealand's Environmental Reporting Series

2014 Air domain report

DATA TO 2012

newzealand.govt.nz



Environment Aotearoa 2015

DATA TO 2013

newzealand.govt.nz

Purpose of the Environmental Reporting Act

Regular environmental reports

Independent production of environmental reports

Advice to decision-makers is based on **accurate** and **credible evidence**

Move the public conversation **away from debating the data** towards addressing environmental issues

Fair and **accurate** environmental reporting

Environmental reports are **trusted by the public**

Certainty about what will be in environmental reports

Framework

- Pressure-state-impact analytical framework.
- Air, atmosphere and climate, freshwater, land, marine, and biodiversity domains.
- Topics, set by Ministers.
  - Measurable, environmentally significant, causally linked.
- Indicators, approved by Government Statistician.
  - Relevant, accurate, timely, accessible, interpretable, consistent/coherent.

Government Statistician's role

- to follow best practice principles and protocols.
- be satisfied that the statistics accurately represent the topic they purport to measure.
- has the sole responsibility for deciding the procedures and methods that are to be used in providing the statistics.
- in producing and publishing an environmental report, the Secretary and the Government Statistician must act independently of any Minister of the Crown.

Quality assurance

- Data quality is a multi-dimensional concept that, when met, generally defines data as ‘fit for purpose’: more than just accuracy.
- NSOs face challenges in conveying data quality rather than omitting the measure, unless these limitations would result in misuse of the data regardless of any caveats that are put in place.
- Quality challenges permeate the general statistical business process model.

- Unique range of data sources, collected for various reasons, not always to official statistics standards.
- Quality judgements are a result of holistic decisions based on:
  - Uses
  - Costs
  - Conditions and circumstances affecting quality
  - User expectations
- Trade offs between accuracy and relevance.

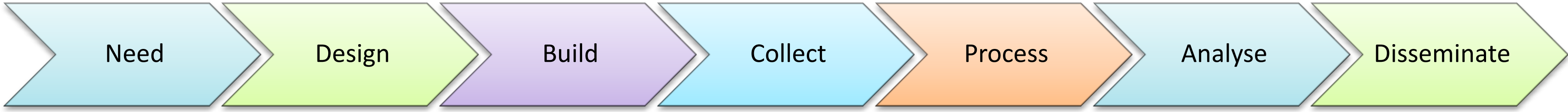
Environment Aotearoa 2015

- The QA process followed two general stages:
  1. A conceptually-focused quality assessment which occurred prior to data collection.
  2. Data checking: ensuring that the actual data was compiled and prepared to the standard expected, and checking for missing values, outliers, unusual movements or levels completed once data has been received.

Testing the Data Quality Framework against environmental data: Process

Conceptual fit checks

Quality assurance of data



Need

- Identified measure may be under-developed

Design

- Measures template prepared – may be incomplete
- Lack of internationally standard methodologies can affect statistical design
- Procurement written with ‘incomplete information’

Build

- Collections are undertaken by a variety of agencies, supplied in various forms – unclear what we are expecting to receive
- Multiple versions may be supplied

Collect

- Vast majority of data are collected from external (official and non-official) agencies
- Need to check procured reports and data, and see what metadata gaps remain.

Process

- Large number of input datasets means a range of data quality checks and techniques need to be applied, but done so to a consistent standard
- Need to ensure data received matches expectations.

Analyse

- Generating aggregating outputs from non-random sampled data is not possible – need to understand the data before analysis.
- Lack of coherent analysis tools can impact on validating data quality
- Lack of access to unit record data affects QA and analysis

Disseminate

- Conveying data quality and limitations appropriately; expressing non-representative data in a national context

Criteria	Standard descriptor	Applicability to environment
Relevance	The degree to which the statistical product meets user needs in coverage, content and detail.	-Geographic coverage -Fit to topic -Collection: How long, where from, by who and what for
Accuracy	The degree to which the information correctly describes the phenomena it was designed to measure.	-Accuracy in relation to topic -Methods and limitations -Available metadata
Timeliness	The degree to which data produced are up-to-date, published frequently and delivered to schedule.	Five years or less for key statistics
Accessibility	The ease with which users are able to **access** and understand the statistical data and its supporting information.	Extensive use of modelling -Transparency -Underlying data -Peer review
Coherence /consistency	The degree to which statistical information can be successfully brought together with other statistical information within a broad analytical framework and over time.	-Comparability with similar international indicators -Coherency across measures -Time-series consistency
Interpretability	The availability of supplementary information and metadata necessary to interpret and **use** the statistics effectively.	Ease by which a user can understand/track how the raw data feeds into the indicator.

Key issues

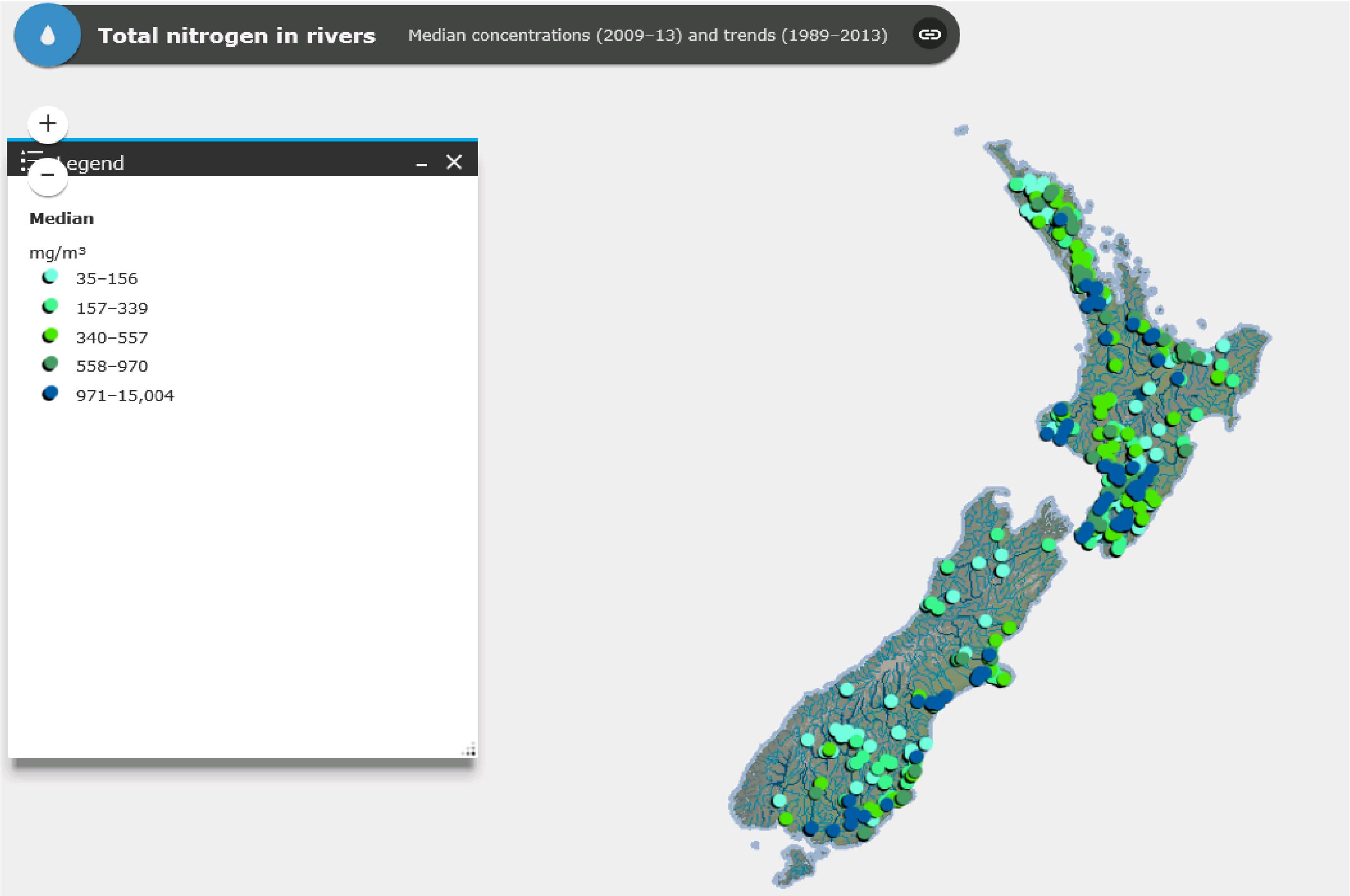
- Frameworks and relevance
- Representativeness and aggregation
- Timeliness
- Coherency
- Accessibility

Relevance

- Relevance can be expressed in relation to the:
  1. conceptual or statistical framework;
  2. use in national-level reporting;
  3. topic completeness (ie includes all required information at any given point in time for a given location).
- Linking indicators in the DPSIR indicator framework is problematic:
  a) other uncontrolled factors affect the relationship between indicators;
  b) compiled using different classification systems, or
  c) are not linked on a spatial or temporal basis.

Representativeness and aggregation

- Many environmental measurements are at ‘problem areas’ and unrepresentative, leading to inherent (upward) bias in any aggregate.
- As the sample is non-random, statistical tests on that sample cannot be undertaken.
- Data quality issues mean that national level point-estimates can therefore not be derived.
- Clarity required on what ‘national-level reporting’ means.



Total nitrogen in rivers Median concentrations (2009–13) and trends (1989–2013)

Legend

Median

mg/m³

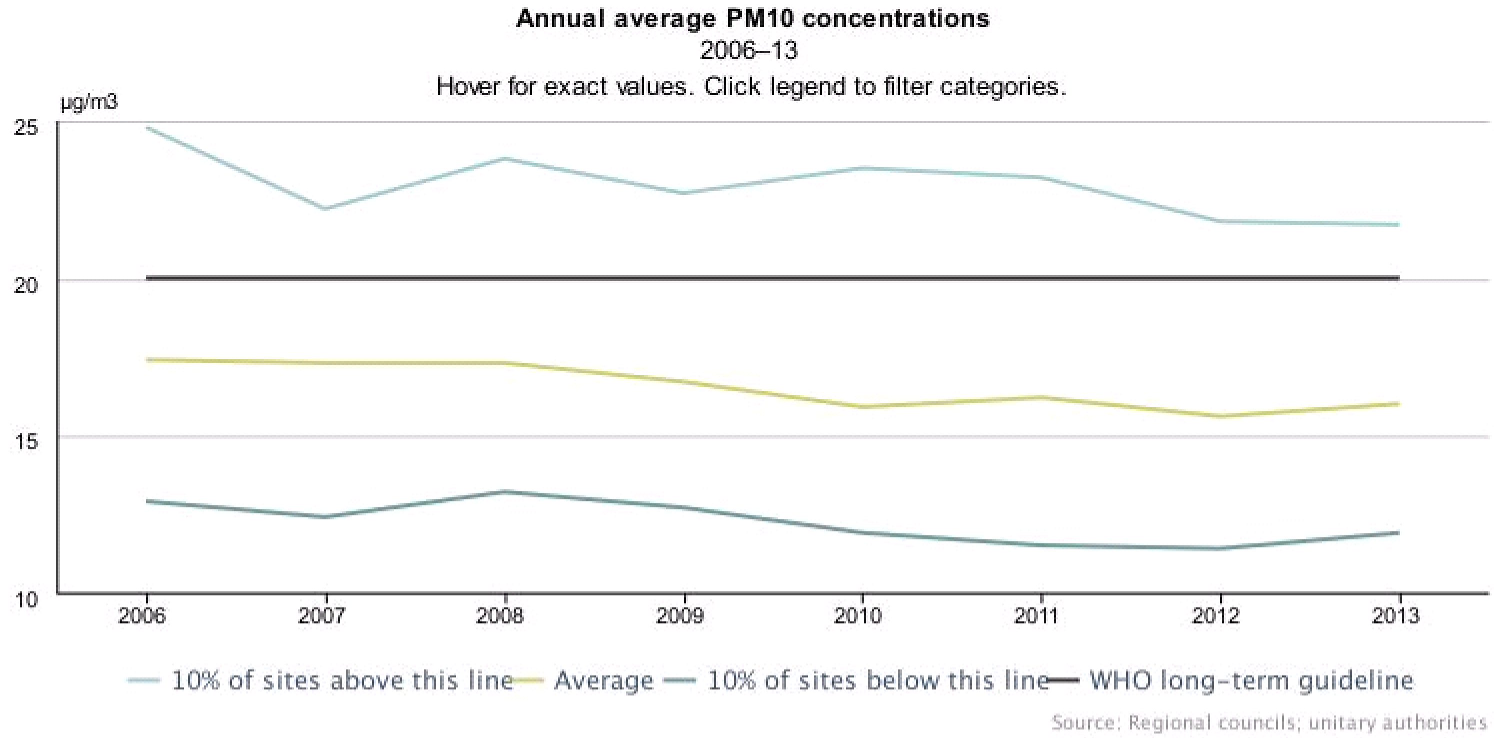
● 35–156

● 157–339

● 340–557

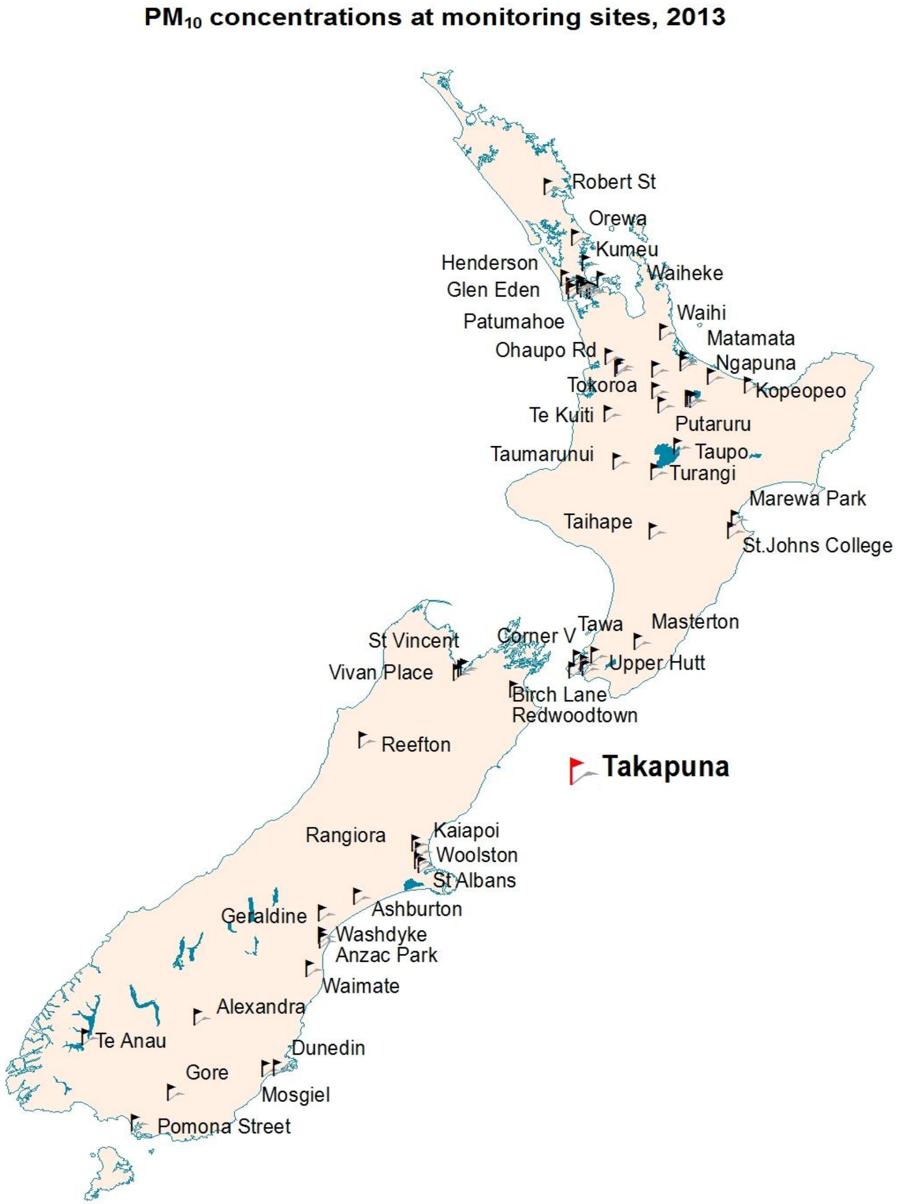
● 558–970

● 971–15,004



Source: Regional councils; unitary authorities

- *Timeliness* – many datasets infrequently collected, many used to derive other measures but various versions used.
- *Coherency* – time series: many reporting changes to account for; international: need accepted international methodologies.
- *Accessibility* – often only have aggregated data to analyse, cannot assess sub-annual or lower level data, or appropriateness of transformations.



The coordinates given for Takapuna were 1756059, 5**29**8077 but should have been 1756059, 5**92**8077

Summary

- Clear need for NSOs to ensure environmental data quality is assessed.
- NSOs have a comparative advantage through experience in using data quality frameworks.
- Environmental statistics raises challenges for, but does not negate the applicability of, standard data quality frameworks.

Starter questions for discussion

- Do these issues resonate with those of other countries?
- What other key data quality issues have statisticians encountered?
- How can data quality considerations become more fully integrated into the production of environmental statistics?