Dealing with large particle counting data sets

A.D. Ceronio and J. Haarhoff

Rand Afrikaans University, PO Box 524, Aucklandpark, 2006, South Africa (E-mail: *jh@ing.rau.ac.za*)

Abstract A particle count survey of 21 South African water treatment plants over a period of 15 months presented the authors with challenges similar to those experienced by other workers in the field. The amount of data generated was staggering and had to be dealt with in an orderly and structured way to derive the maximum benefit from the survey. Furthermore, the number of significant data points involved if the entire count is to be taken into consideration complicated interpretation of particle counting data. This led to the application of several data reduction techniques to reduce the number of significant parameters that had to be considered during analysis. The most common conventional method is the use of the total particle count larger than a given size, for example total count per ml > 2 μm. This method, however, negates one of the most powerful abilities of the particle counter, namely the ability to indicate particle size distribution. The application of the power law, a common alternative, provides a more detailed description but has its flaws. In this paper the authors illustrate how many particle counts were successfully handled in a purpose-designed database and how the power law concept was improved to provide a better particle counting data-reduction methodology.

Keywords Data management; data reduction; particle counting; power law

Water Science and Technology: Water Supply **Vol 2 No 5–6 pp 35–40**

Introduction

A survey of South African water treatment plants was undertaken to determine reference levels of particle counts in the raw and final waters as well as the intermediate process steps. In all, 21 treatment plants were surveyed in a period of 15 months. The plants were monitored on at least three different days, each monitoring session including the complete treatment profile from raw to final product. Sixteen particle counting bin sizes were set, but in the end thirteen bins sizes were used for further analysis, covering a size range between 2 μm and 50 μm. Many thousands of particle counts were generated and the need for systematic data handling and significant data reduction was clear; more so if it is considered that a direct particle count is difficult to interpret and a substantial amount of post-measurement data manipulation is required to determine parameters such as normalised counts, particle volume fractions and volume average particle sizes among others.

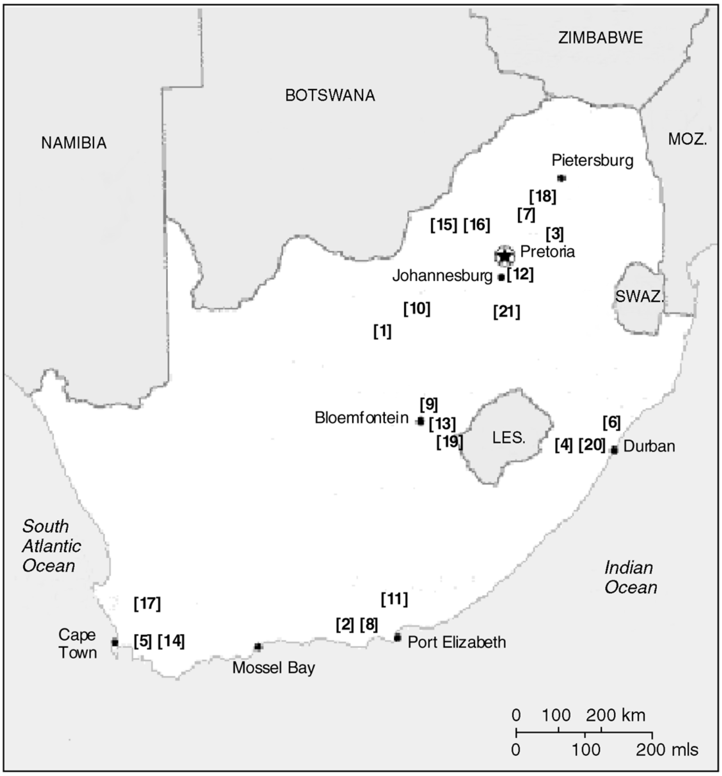
This paper expands on the structure of the database used to assist in the handling and post-processing of the data as well as the improvements achieved in the area of data reduction.

The survey

The national survey of treatment plants (Figure 1) was planned and executed in such a way that the most important raw water types and most prominent treatment processes would be covered. The raw waters included low turbidity waters, eutrophic waters, highly turbid waters and coloured waters. The major process steps monitored included coagulation/flocculation, vertical settling, clariflocculation, horizontal settling and filtration. Other steps monitored included oxidation, lime addition and flotation.

The database

In all 108 plant visits were made and 192 individual process runs were monitored that resulted in 1386 separate and final process data points. Each process data point is the



[1]	Balkfontein	[8]	Loerie	[15]	Vaalkop, Plant 2
[2]	Churchill	[9]	Maselspoort	[16]	Vaalkop, Plant 3
[3]	Cullinan	[10]	Midvaal	[17]	Voëlvlei
[4]	Durban Heights	[11]	Nooitgedacht	[18]	Wallmannsthall
[5]	Faure	[12]	Rietvlei	[19]	Welbedacht
[6]	Hazelmere	[13]	Rustfontein	[20]	Wiggins
[7]	Klipdrift	[14]	Steenbras	[21]	Zuikerbosch

Figure 1 Locations of the plants covered in the survey

median of 3 to 7 separate counts. The processing and organisation of the large number of data points was simplified due to the use of a database (Figure 2). Each count was read into the database after the data were first labelled with important information regarding the count and any qualifications on the data were stated. The data were read into the database in cumulative form. Other measurements taken at the time of the particle count (turbidity, pH, conductivity, suspended solids, temperature and zeta potential) are also included in the database. Provision is made for the direct importation of data into the database if required.

Once read into the database the following calculations are done automatically from the cumulative count data, thereby avoiding the tedious task of calculating the parameters for each count:

- differential counts
- normalised counts
- determination of the power law regression coefficients
- incremental particle volume contributions based on measured data
- total particle volume based on measured data
- volumetric average particle diameter based on measured data

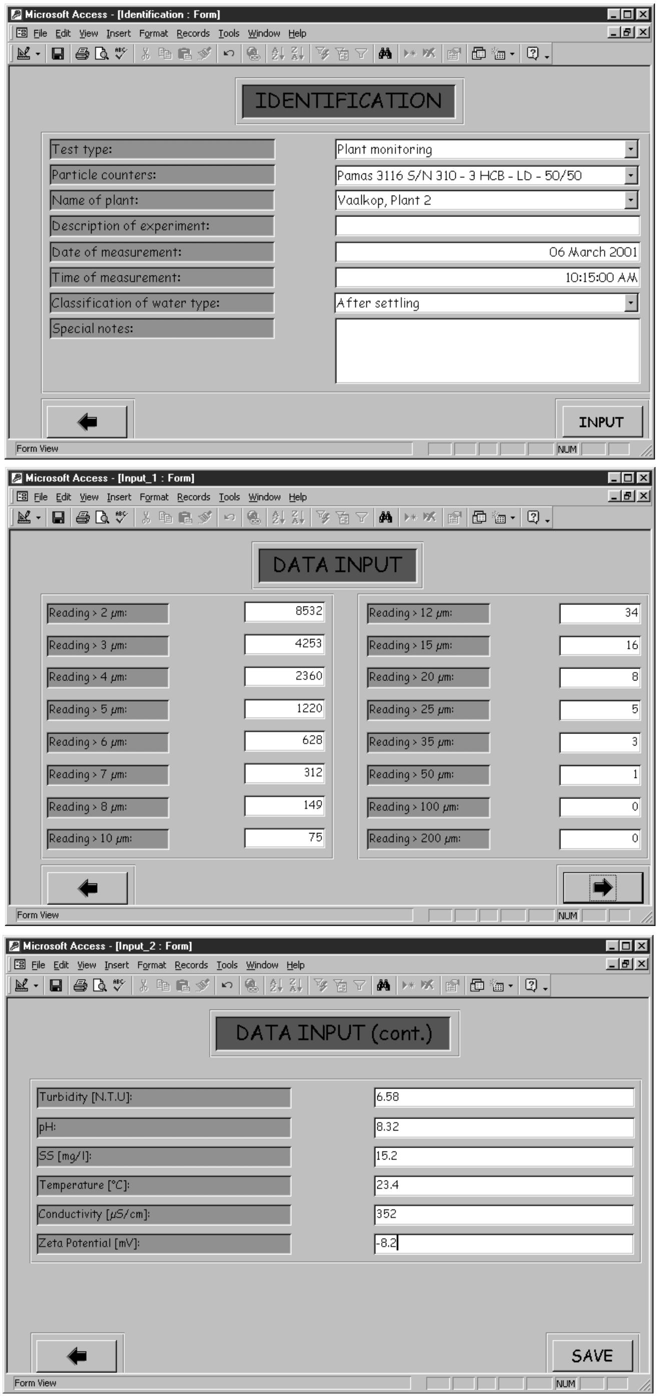


Figure 2 Particle counting database input screens

- smoothed normalised particle counts calculated from the power law coefficients
- smoothed incremental particle volume contributions calculated from the power law coefficients
- smoothed total particle volumes calculated from the power law coefficients
- smoothed volumetric average particle diameter calculated from the power law coefficients

The database employed was a commercial product (Microsoft Access) and the associated programming was done using Visual Basic. Once the 30 required parameters per sample have been entered, an additional 125 cells of data are generated.

The calculated values are automatically stored on the database and are immediately accessible. The use of the database has greatly enhanced the availability and accessibility of data in this project.

Data reduction

Data reduction on particle counting data is required to reduce the number of parameters that have to be considered during data evaluation and also to smooth the counts. A preliminary analysis of the data set at this points confirmed a well known problem inherent to particle counting, namely that the counts drop off so rapidly with increasing size that the counts in the large size bins are very low and therefore erratic. For particle volume calculation, however, these erratic "large-bin" counts dominate and therefore skew the result. It is therefore necessary to find and fit a statistical distribution through the counts to allow more reliable estimation of the counts in the larger size bins.

The most common model used in the literature is the power law, which reduces the normalised (n_n) count in the i-th bin to a two-parameter distribution (A, β) with d being the geometric mean size describing the bin under consideration.

$$n_{ni} = A.d_i^{\beta}$$

This distribution is convenient, since it plots as a straight line on a log-log plot and regression is rapidly done with standard textbook methods. Fundamentally, however, the power law is flawed. Lawler (1997) showed that b couldn't be accepted as constant, as this would extrapolate to an infinite particle number at very small sizes as well as an infinite particle volume at very large sizes. He suggested a variable-b model as alternative, which allows b to vary with particle size, leading to another two-parameter distribution that would be consistent with zero particle counts for small particles and zero particle volume for large particles:

$$n_{ni} = A.d_i^{a \log d_i} = 10^{\alpha} d_i^{a \log d_i}$$

This study developed Lawler's suggestion into a workable model. Taking the least sum of squares (SS) approach, values can be found for A and a that would approximate the measured normalised values (n_n) with predicted normalised values (n_n'). This is done for the 13 bins used between 2 and 50 µm:

$$SS = \sum_{\text{Bin}=1}^{13} [\log(n_n) - \log(n_n')]^2$$

$$SS = \sum_{\text{Bin}=1}^{13} [\log(n_n) - \log(A) - (a.\log(d_g))^2]^2$$

The unique values for A and a that describe a specific particle count are then determined by solving the following two equations:

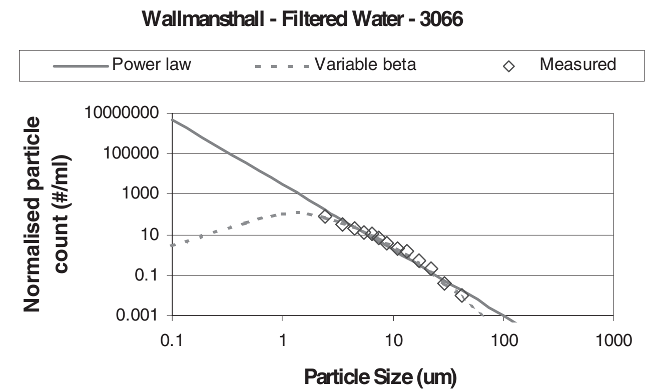
$$\frac{\partial SS}{\partial A} = 0$$

$$\frac{\partial SS}{\partial a} = 0$$

The indices α (=log(A)) and a, respectively, are then used to describe the particle count and size distribution in place of α and β derived from the power law. The index α has the same meaning in the two models, that is the log of the normalised particle count between 0.5 µm and 1.5 µm, although the two models will generate different values for this index, and the indices a and β describe the particle size distribution.

Figure 3 contains a comparison of the two models in terms of normalised particle count and normalised volume contributions. The graphs illustrate the more typical distribution where the measured counts diverge from the linear correlation as assumed by the power law. The variable-β model does not allow for independent manipulation of the curvature of the model. Although this is feasible, it would imply the inclusion of a third variable. The modelling of the normalised volume contributions is also significantly more accurately done using the variable-β model (Figure 3). Figure 4 illustrates a frequency plot of correlations for the power law and the variable-β model. This plot uses data derived from more that 1400 individual counts performed on a wide range of water types ranging from raw water to final effluent and it includes all the intermediary process steps monitored in the survey of plants. Figure 4 therefore illustrates the improved descriptive capabilities of the variable-β model over the power law.

In general, the variable-β model provided significantly better fits than the power law and the variable-β model was therefore selected as a reliable representation of particle counts, further reducing the thirteen data points per sample to two parameters. The database was finally adapted to automatically integrate the variable-β model to provide reliable particle



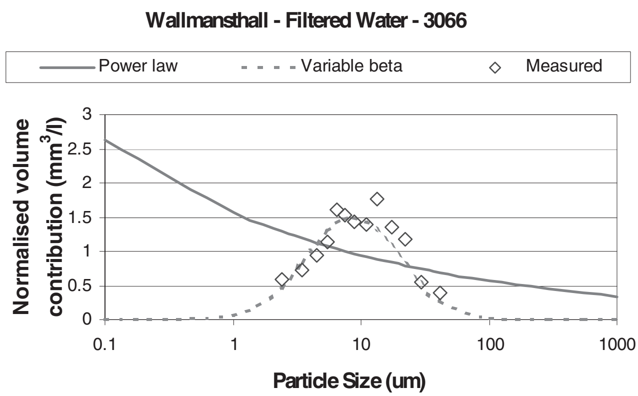


Figure 3 Comparison of the power law and the variable-β models on two random particle counts in terms of normalised particle count and volume contribution

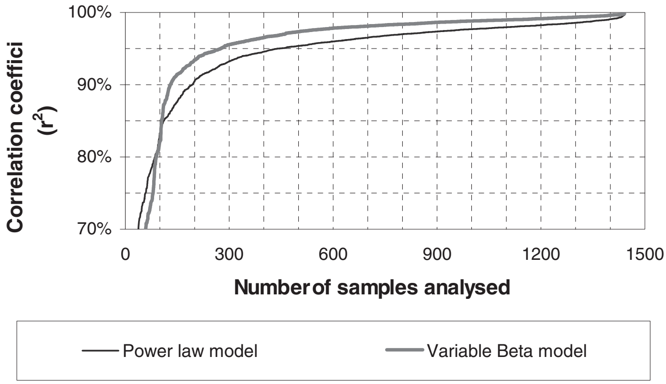


Figure 4 Frequency plot of correlations (r^2) of the power law and the variable β model to measured data

counts and particle volumes for further analysis. The following capabilities were programmed into the database:

- determination of the variable-β regression coefficients
- smoothed normalised particle counts calculated from the variable-β coefficients
- smoothed incremental particle volume contributions calculated from the variable-β coefficients
- smoothed total particle volumes calculated from the variable-β coefficients
- smoothed volumetric average particle diameter calculated from the variable-β coefficients.

Conclusions

The following conclusions are drawn from this paper.

- Particle counting data requires a substantial amount of post-processing in order to unlock the full potential of the data.
- If large datasets are generated, a structured approach is necessary in data management.
- Databases provide both the ability to store the data in an orderly fashion and can be designed to take care of the tedious function of post-measurement processing allowing direct access to more meaningful parameters.
- The variable-β model provides an improved description of particle size distribution.

Acknowledgements

The authors acknowledge the financial support of the South African Water Research Commission (WRC) in sponsoring a portion of the work undertaken in this paper.

Reference

Lawler, D. (1997). Particle size distribution in treatment processes: theory and tractice. *Wat. Sci. Tech.* **36**(4), 15–23.