Journal of Mathematical Modelling and Algorithms (2005) **4:** 289–305
DOI: 10.1007/s10852-005-9005-2

Uncertainty Evaluation in Multivariate Analysis – A Test Case Study

RAGNE EMARDSON[1,★], PER JARLEMARK[1] and PER FLOBERG[2]
[1]*SP, Swedish National Testing and Research Institute, Borås, Sweden.*
e-mail: ragne.emardson@sp.se
[2]*SIK, The Swedish Institute for Food and Biotechnology, Göteborg, Sweden.*

Abstract. We have used different multivariate analysis methods to estimate quantities in the fields of food control and atmospheric remote sensing. In order to estimate the uncertainties in these estimates we studied analytical as well as non-parametric numerical methods. The methods have been evaluated by comparison between obtained results and independent sets of measurements. We present one test case from each field, including results, where these methods have been applied. For the food control test case reduced chi-squared (χ^2_ν) of approximately unity indicate that both the analytical and numerical methods used for uncertainty estimation produce uncertainties of reasonable size. In the atmospheric remote sensing test case, a $\chi^2_\nu = 46$ indicated that the uncertainties from the numerical method were far too small, whereas a $\chi^2_\nu = 1.5$ indicate that the size of the analytically determined uncertainties can represent the size of the "true" errors.

Mathematics Subject Classifications (2000): 62H12

Key words: multivariate analysis, uncertainty.

1. Introduction

Measurements abound in such diverse fields as process control, satellite navigation and remote sensing, food control, and pharmaceutical development. The usefulness of these measurement results are very much dependent on the quality of the uncertainty estimates that accompany them. The term multivariate analysis is used for a group of methods for investigating large data sets with many variables. These methods are often well suited for problems where several variables are highly correlated, which otherwise can cause problems if the correlation is not fully known. We have studied both analytical and non-parametric numerical methods for uncertainty estimation in multivariate analysis. The methods have been evaluated by comparison between obtained results and independent sets of measurements. In this paper, we present two different test cases from the fields of food processing and remote sensing, including results, where these methods have been applied.

★ Corresponding author.

2. Regression Methods for Static Processes

The technique to relate two sets of parameters X and Y, where X is measured and Y is sought, is commonly referred to as multivariate analysis. Usually, the goal is to find a relationship between a sought property that is very time consuming to obtain and predictor variables that are relatively easy to measure. If the relation is relatively linear, this can be expressed as

$$Y = X\,B + \epsilon \tag{1}$$

where B contains the coefficients relating X to Y and ϵ contains the model residuals. By calibrating the system with known parameters X and Y, B can be estimated. This estimate can then be used to predict new Y_0 based on new measurements X_0. A straightforward method to estimate B is to solve for the variable using a regular least-squares formulation. However, in many applications, this method may fail due to collinearity in X. Methods exist to deal with this calibration problem. Two commonly used methods are principle component regression (PCR) and partial least squares (PLS). In PCR, the main purpose is to express X by using fewer variables and thereby reduce the sensitivity to noise. In PLS, latent variables of X are calculated in order to maximize the correlation between X and Y. In PCR, the reduction of variables is made after rewriting the relationship between X and Y using two equations as

$$Y = T\,C + \epsilon_y \tag{2}$$

$$X = T\,P + \epsilon_x \tag{3}$$

where P is chosen as the eigenvectors of the covariance matrix of X and T is a reduced set of modified measurements. A corresponding rewriting is performed for the PLS method.

For a broader presentation of multivariate techniques, including methods based on maximum likelihood parameter estimation methods, see, for example, Höskuldsson [5], Wentzell et al. [9] and Sundberg [8].

3. Uncertainty Evaluation

Uncertainty measures are important in order to evaluate the significance of derived estimates. We have studied both analytical and numerical methods for uncertainty determination. The uncertainties are specified as variances or standard deviations. Below, we describe the analytical and numerical approaches used. For more on both analytically derived uncertainties and resampling techniques for calculation of uncertainties in multivariate regression coefficients, see, for example, Faber [4].

3.1. PROBLEM IDENTIFICATION

In order to evaluate uncertainties in least-squares estimates, we can use the following formalism. The ideal linear relationship, i.e., the relationship between X and Y if they were measured without errors, is

$$Y = X\,B + \epsilon \tag{4}$$

As measurement errors always exist, we can write the measured quantities

$$\begin{aligned} y &= Y + \nu_y \\ x &= X + \nu_x \end{aligned} \tag{5}$$

A calibrated model, possibly after transformation, can be expressed as:

$$\hat{B} = \left(x^T x\right)^{-1} x^T y \tag{6}$$

and the predicted $\hat{y}_0$

$$\hat{y}_0 = x_0 \hat{B} \tag{7}$$

where as above

$$x_0 = X_0 + \nu_{x\,0} \tag{8}$$

The property of interest is the variance of the predicted values, $Var[Y_0 - \hat{y}_0]$. It can be written using the relationships above as

$$\begin{aligned} Var[Y_0 - \widehat{y}_0] &= Var\left[Y_0 - (X_0 + \nu_{x0})\widehat{B}\right] \\ &= Var\left[Y_0 - (X_0 + \nu_{x0})(x^T x)^{-1} x^T y\right] \\ &= Var\left[Y_0 - (X_0 + \nu_{x0})\left((X + \nu_x)^T (X + \nu_x)\right)^{-1}\right. \\ &\qquad \left.(X + \nu_x)^T \left(Y + \nu_y\right)\right] \end{aligned} \tag{9}$$

This expression is non-trivial, and furthermore, the errors we can use to validate the uncertainty estimates are based on noisy measurements, $Var[y_0 - \hat{y}_0]$. Using these noisy measurements directly often lead to an overestimation of the errors which may lead to belief that the uncertainties are underestimated. The expression for the uncertainties further does not include a time variation in the model B nor systematic deviations from the linear relationship assumed. In the following, we present simplified expressions based on alternative techniques to estimate the uncertainties.

3.2. ANALYTICAL METHODS

An analytical expression for the variance of the predicted values based on the PCR algorithm can be derived (see, e.g., [5, 7]).

$$Var\left[Y_0 - \hat{Y}_0\right] = \sigma_\epsilon^2\left(1 + t_0\left(T^T T\right)^{-1} t_0^T\right) \tag{10}$$

where σ_ϵ^2 is the variance of the model residuals and T and t_0 are defined by Equations (3) and (8). This relation is based on assumptions that all errors can be attributed to the model, C in Equation (2). This is not a fully realistic assumption, as described above, and more elaborate expression for uncertainties have been derived by, e.g., Hoy et al. [6]. Corresponding expressions for PLS would be very similar to that of PCR and are therefore not included here.

3.3. NUMERICAL METHODS

Non-parametric estimation methods of statistical errors have the attractive property of requiring very little modeling or assumptions. Two of the most frequently used non-parametric methods in univariate analysis are the bootstrap and jackknife techniques. Both these techniques are based on resampling and are described in Efron and Gong [2].

In our evaluations, bootstrap and jackknife techniques tended to underestimate the errors in multivariate analysis. We developed a hybrid method that takes into account the correlation in X. The basic principle behind the hybrid method is to generate independent subsets of X and Y and from these derive $\hat{C}$ or $\hat{B}$ for the statistical computation. By incorporating these data sets in the computation of the variance, the uncertainties are less likely to be underestimated as using, for example, jackknifing. Further, the number of samples used in the statistical computation is a trade-off between the number of values used for the statistics and the number of latent variables. All results presenting numerically derived uncertainties below are produced using this hybrid method.

4. Case Studies

We have studied two data sets with the main purpose of assessing the methods to estimate uncertainties. The data sets are from different areas: food analysis and remote sensing. We hereby refer to these as Case 1 and Case 2. In order to evaluate the quality of the estimated uncertainties, we compare these to "real" errors obtained as the difference between predicted and measured values. Statistically, we use the reduced χ^2 (e.g., [1]), which optimally is equal to one if the estimated and "real" errors agree.

4.1. CASE 1: SALMON DATA

Case 1 is based on data from the food-processing industry. Food quality parameters are of interest for both manufacturers and consumers. The data set originates from an EU-funded project with the objective to develop and validate a new rapid method and instrumentation to determine the quality of seafood products. It has been shown within this project that the quality of fresh and frozen fish can be measured with the use of a probe-based, microwave instrument. The instrument will be developed further by a company in Germany.

Table I. Properties used for the study of salmon

Measurements on gutted fresh salmon	
1	Storage days, (+1°C), after slaughter of salmon
2	Demerit points according to QIM, Quality Index Measurement. A standardized sensory evaluation of the fish quality
3	Total length of gutted fresh fish, head to tail (cm)
4	Total weight of gutted fresh fish (kg)
5	Quota "3"/"4" (cm/kg)
6	Quota 100*"4"/"3" (kg/m)
Measurements on minced salmon flesh	
7	Water content according to standard NMKL 23 (g/100 g)
8	Protein according to Mod NMKL no. 6, Kjeltec (g/100 g)
9	Energy calculated (kJ/100 g)
10	Carbohydrates calculated (g/100 g)
11	Ash content NMKL7, 23 (g/100 g)
12	Raw fat, NMKL 131 (g/100 g)
Measurements on ground and heat-treated salmon flesh	
13	Moisture loss during centrifugalizing (%)
Liquid from above moisture loss measurement	
14	Fat removed from moisture loss sample after freezing (%)
15	Dry matter of moisture loss sample after fat removal (%)
16	Water content calculated by moisture loss–fat–dry matter (%)
Color D65/2° measurement of minced salmon flesh measured by Minolta CR-300	
17	L*
18	a*
19	b*
20	C*
21	h°
Similar to above, measured as a difference from a white standard plate	
22	DL*
23	Da*
24	Db*
25	DC*
26	DH*
27	DE*ab
Logarithm of storage days	
28	ln(storage days), "1"

In this study, we measure permittivity, ϵ' and ϵ'', using reflection measurements on 202 different frequencies. The measurements are made on five salmon filets on the flesh side at nine different locations and a repetition on the first location, in total 10 measurements for each filet. This procedure is repeated at 10 different occasions. One of the five salmon filets is used for validation and therefore not in the calibration of the model. Hence, the variable X for calibration contains the measured permittivity and is of size 400×202. That is, X contains measurements of 202 variables measured on four samples at 10 positions at 10 distinct occasions.

Table I shows the seafood product quality parameters sought in this study. In total, there are 28 properties. The variable Y containing the measurands is hence of size 400×28. That is, Y consists of the 28 quality properties we try to estimate of which we have 400 measurements each.

Figure 1 shows the estimated value of each of the 28 Y properties for one example. These properties are estimated for the salmon filet chosen for validation using PCR. The example estimate corresponds to a measurement at one single location at one occasion. With each presented estimate, two uncertainty values are associated. These are analytical and numerical uncertainties respectively. Values range from relatively large values for the length (variable 3) and water content (variable 7) to small values for the calculated carbohydrates. Also, the uncertainties for the different variable vary significantly. Figure 2 shows the

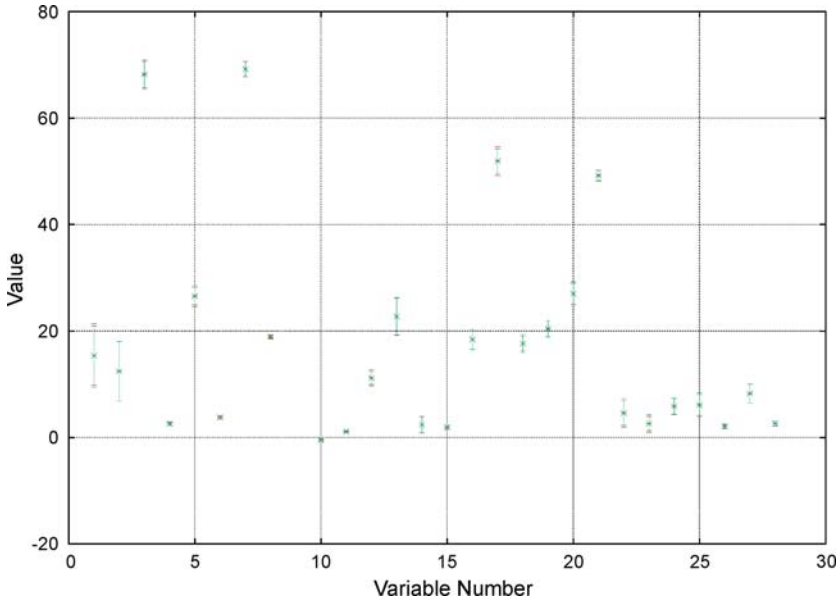

Figure 1. Estimated value of each of the 28 Y variables of Test Case 1 using the existing data set. With each value, two uncertainty values are associated: the analytical (red) and numerical (green) uncertainties, respectively.

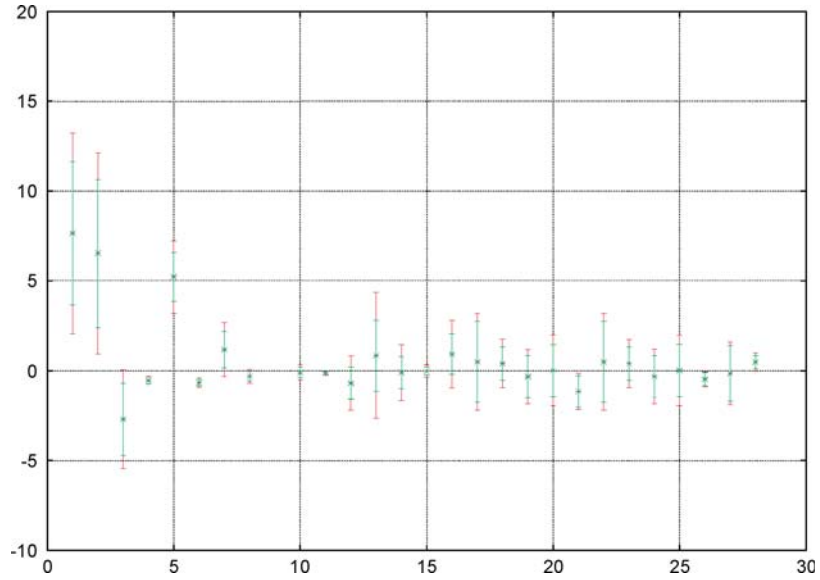

Figure 2. Estimated value of each of the 28 Y variables of Test Case 1 with the "true" reference values withdrawn. The analytical (red) and numerical (green) uncertainties are shown as error bars.

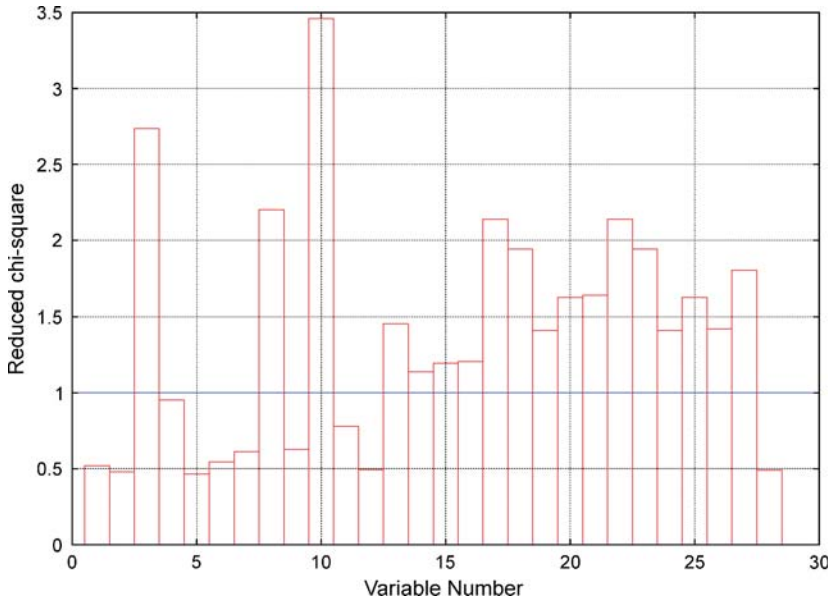

Figure 3. Reduced chi-squared of the deviations of the estimates of Test Case 1 from the "true" reference values. The analytical uncertainties have been used in the calculations.

estimated value with the "true" values withdrawn. Again, each value has two associated uncertainties, analytically and numerically determined.

We have calculated χ^2_ν for all 28 properties based on the salmon filet chosen for validation. Each χ^2_ν value is based on all 100 measurements on this filet. Figure 3 shows χ^2_ν for the analytically determined uncertainties. Figure 4 shows corresponding χ^2_ν for the numerically determined uncertainties. According to the figures, both the analytical and numerical methods have difficulties determining the uncertainty for the carbohydrates (variable number 10). Also, the length (variable 3) and raw protein (variable 8) uncertainties are underestimated. Overall, most χ^2_ν values are between 0.5 and 2 for both techniques. Please note the different scale in the two figures displaying the chi-squared values.

4.2. CASE 2: ATMOSPHERIC MICROWAVE RADIOMETRY

Test Case 2 is based on data from a microwave radiometer sensing the atmospheric radiation at 21.0 and 31.4 GHz. The data are used to estimate the amounts of atmospheric water vapor and liquid water in a line of sight. A major application of the water vapor estimates has been the assessment of global positioning system (GPS) measurements. The parameter of interest is then the equivalent zenith wet delay (ZWD), the apparent excess radio propagation path due to water vapor. The estimated cloud liquid water, represented by the zenith amount of liquid water (ZLW), is a quantity mainly of meteorological interest.

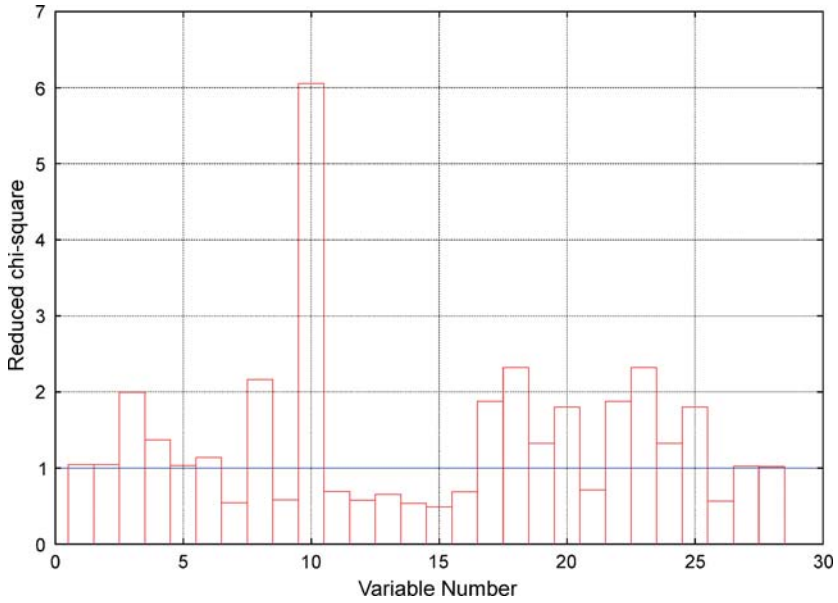

Figure 4. Reduced chi-squared of the deviations of the estimates of Test Case 1 from the "true" reference values. The numerical uncertainties have been used in the calculations.

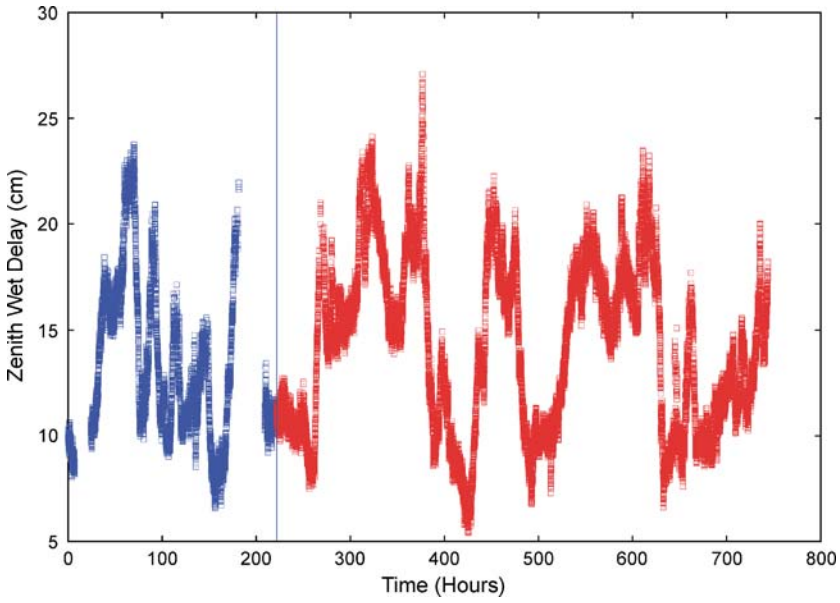

Figure 5. The reference zenith wet delay data used in Test Case 2. The left and right curves represent the part of the data set that were used for calibration and verification, respectively.

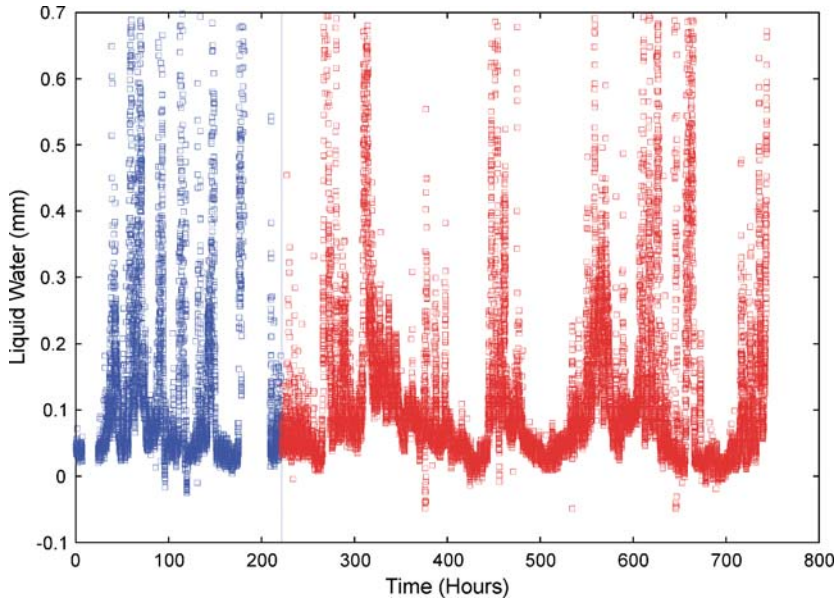

Figure 6. The reference zenith liquid water data used in Test Case 2. The left and right curves represent the part of the data set that were used for calibration and verification, respectively.

Instrumental calibration is required to relate the measured radiometer detector voltages to the sought quantities ZWD and ZLW. The normal calibration procedure consists of observations of the atmosphere in a multitude of directions and therefore demand mechanic steering of the radiometer, see, e.g., Elgered and Jarlemark [3]. Radiometers that rely on this way of calibration are in general expensive to construct and hence not widely spread globally. Calibration procedures not requiring mechanic steering would enable a simplified instrument design.

We have investigated the possibility to use multivariate techniques in order to calibrate radiometers without mechanic steering. For this study, we have used a data set consisting of 1 month of radiometer data, the month of August 2001, all from one direction on the sky. The reference values of ZWD and ZLW, estimated using the normal calibration procedure, are given in Figures 5 and 6. We used the first quarter of the ZWD and ZLW data sets as a calibration measurand Y. We assumed that these data could be found by, e.g., a visiting steerable radiometer. The principle measurements of the radiometer, detector voltages at 21.0 and 31.4 GHz, for the month are found in Figures 7 and 8. In order to aid the calibration with local meteorological data, we used the pressure, temperature, and relative humidity (see Figures 9–11). The detector voltages, pressure, temperature, and humidity of the first quarter of the data sets then form X in the multivariate calibrations.

The PLS method was used to calibrate the system. We performed verification on the ZWD only as this is the premium observable for such radiometers as

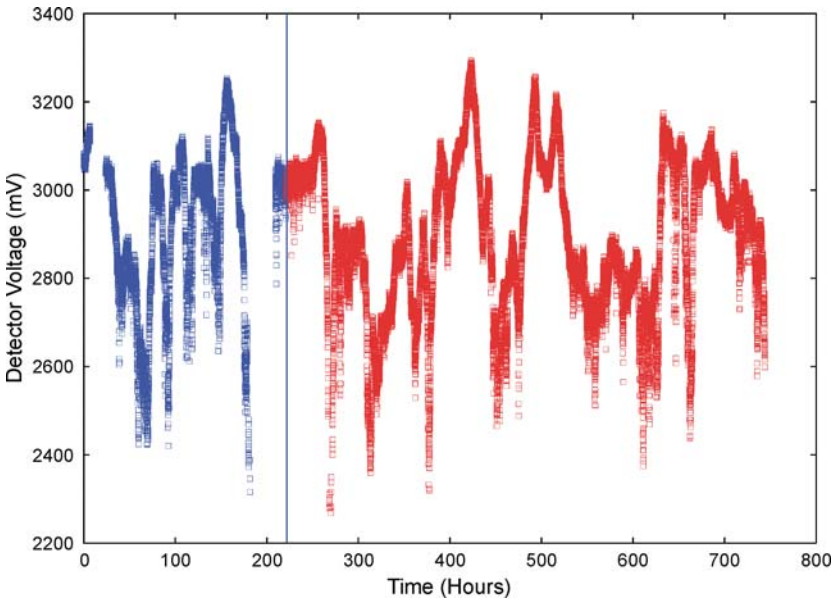

Figure 7. The 21.0-GHz radiometer detector voltage for the data set of Test Case 2. The left and right curves represent the part of the data set that were used for calibration and verification, respectively.

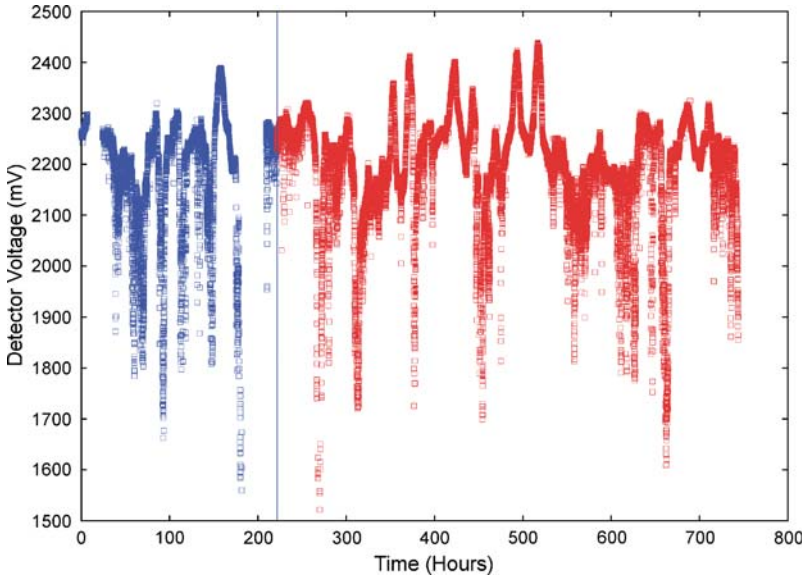

Figure 8. The 31.4-GHz radiometer detector voltage for the data set of Test Case 2. The left and right curves represent the part of the data set that were used for calibration and verification, respectively.

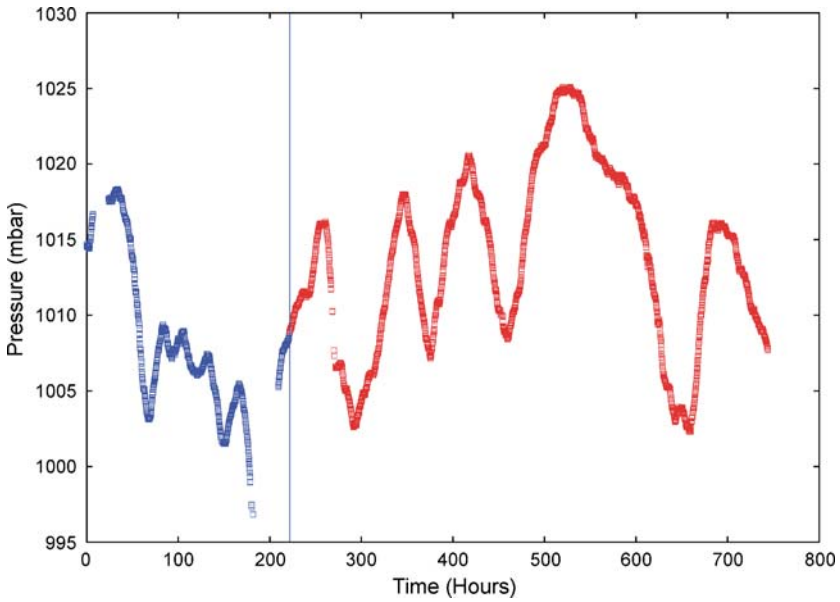

Figure 9. The ground surface pressure for the data set of Test Case 2. The left and right curves represent the part of the data set that were used for calibration and verification, respectively.

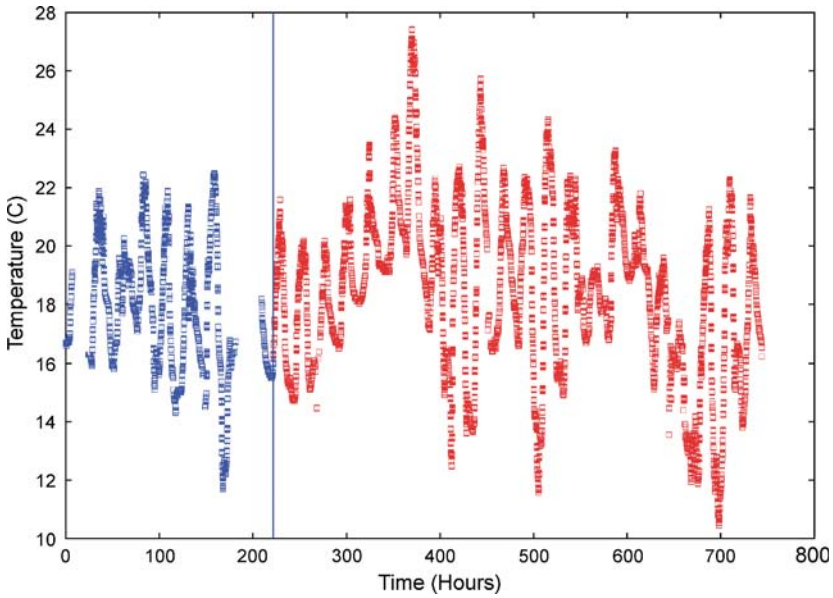

Figure 10. The ground surface temperature for the data set of Test Case 2. The left and right curves represent the part of the data set that were used for calibration and verification, respectively.

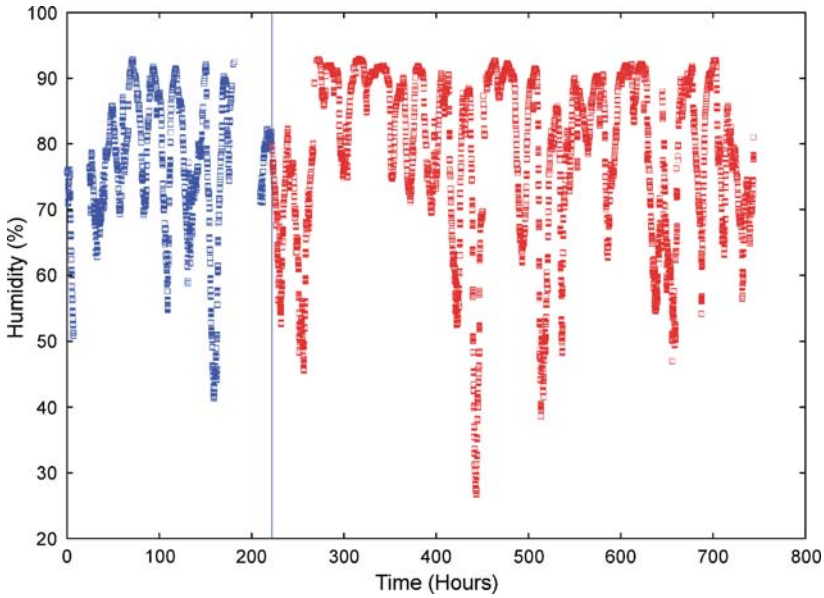

Figure 11. The ground surface relative humidity for the data set of Test Case 2. The left and right curves represent the part of the data set that were used for calibration and verification, respectively.

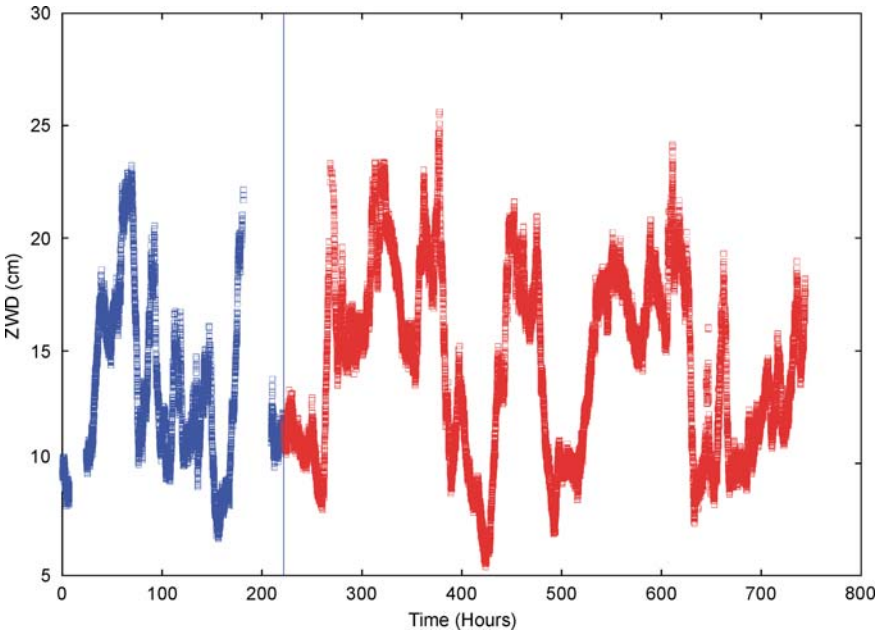

Figure 12. The estimated zenith wet delay of Test Case 2 using PLS for calibration. The left and right curves represent the part of the data set that were used for calibration and verification, respectively.

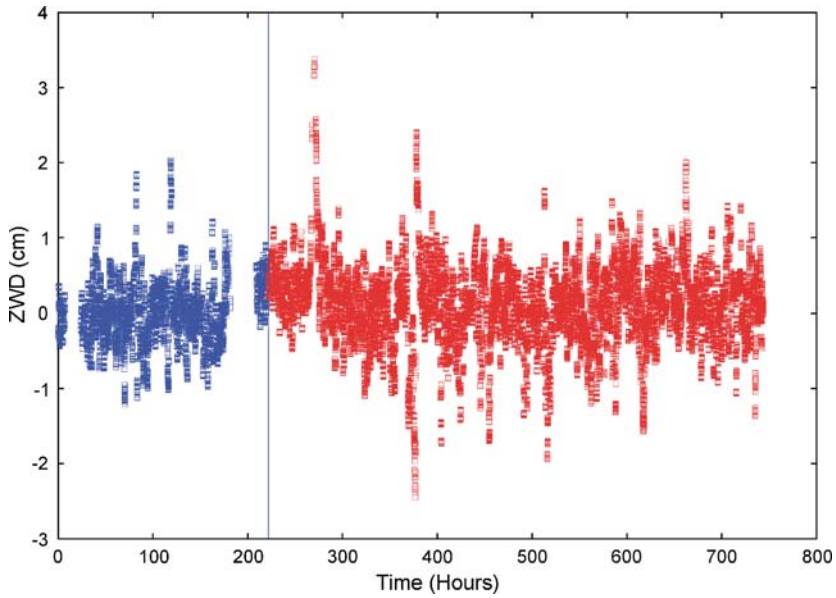

Figure 13. The deviation of the estimated zenith wet delay from the reference data of Test Case 2. The left and right curves represent the part of the data set that were used for calibration and verification, respectively.

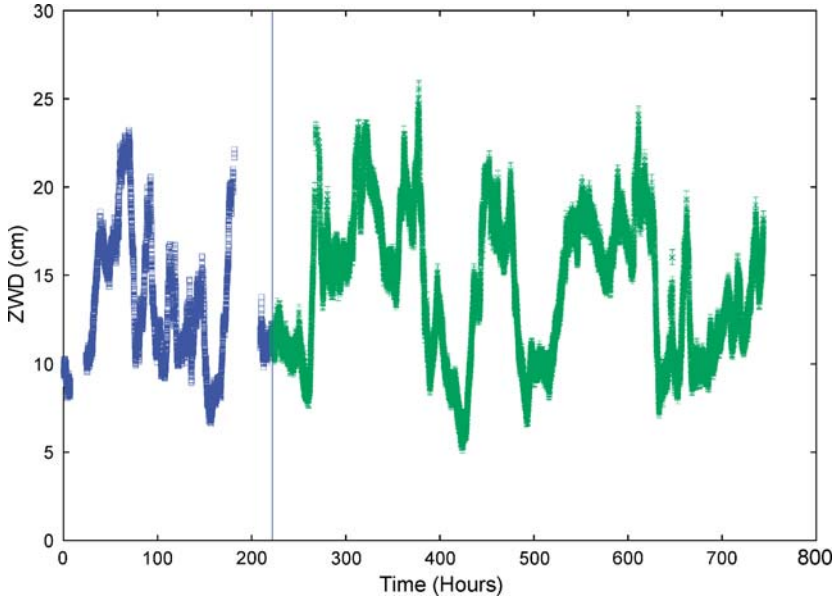

Figure 14. The estimated zenith wet delay of Test Case 2. The left and right curves represent the part of the data set that were used for calibration and verification, respectively. The verification data are equipped with error bars derived using the analytical expression.

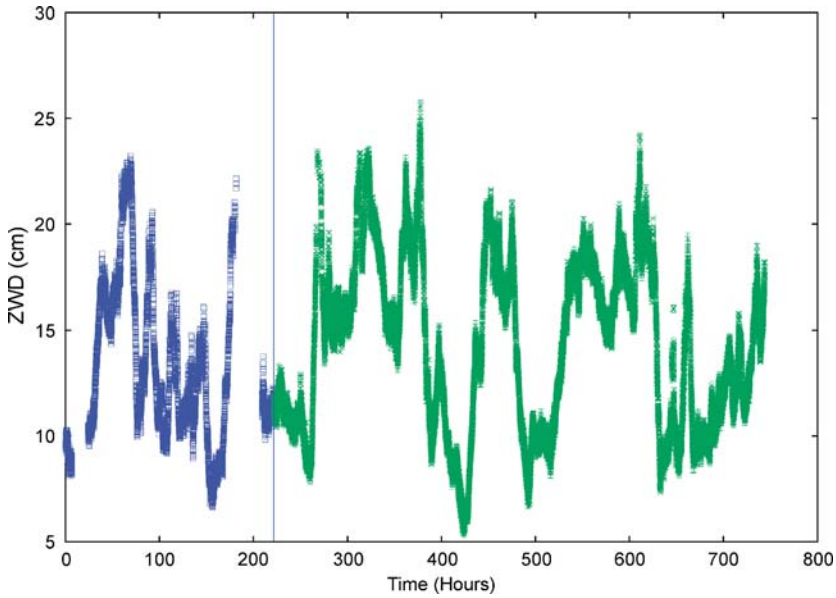

Figure 15. The estimated zenith wet delay of Test Case 2. The left and right curves represent the part of data set that were used for calibration and verification, respectively. The verification data are equipped with error bars derived using the numerical method.

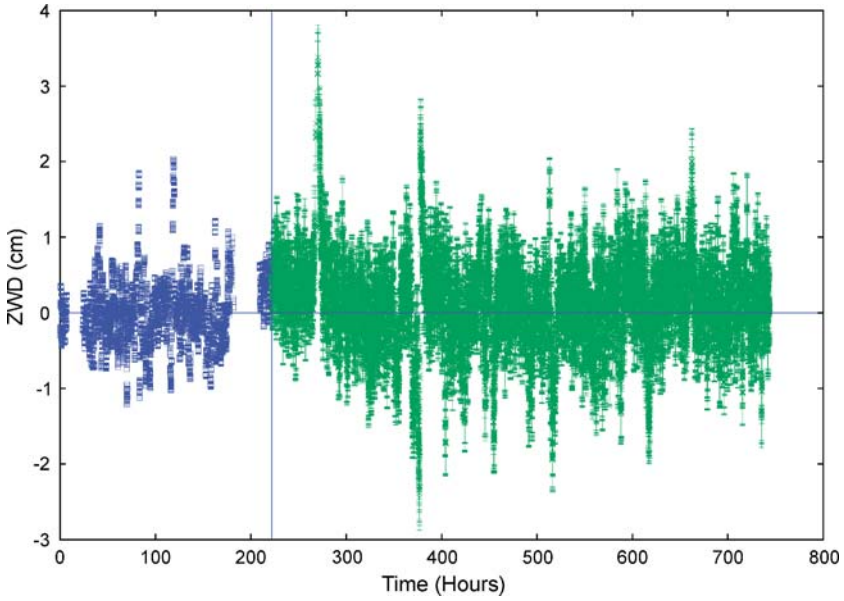

Figure 16. The deviation of the estimated zenith wet delay from the reference data of Test Case 2. The left and right curves represent the part of the data set that were used for calibration and verification, respectively. The verification data are equipped with error bars derived using the analytical expression.

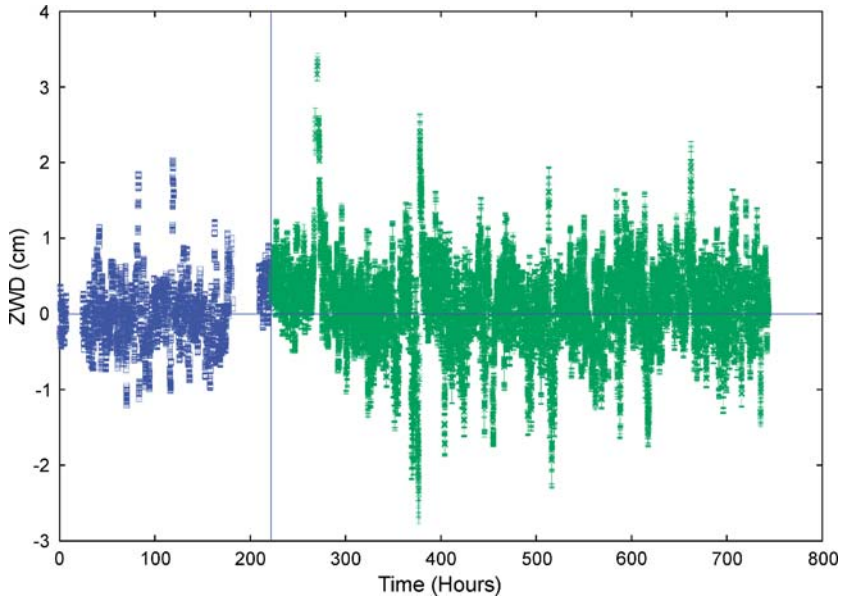

Figure 17. The deviation of the estimated zenith wet delay from the reference data of Test Case 2. The left and right curves represent the part of the data set that were used for calibration and verification, respectively. The verification data are equipped with error bars derived using the numerical method.

stated above. The resulting ZWD estimates are given in Figure 12, and their difference to the reference ZWD values are found in Figure 13. The uncertainties of these estimates, when compared to the reference values, were estimated using the analytical as well as the hybrid numerical method. These uncertainty measures are used as error bars for the estimates in Figures 14 (analytical) and 15 (numerical). These error bars are also applied to the deviations from the reference ZWD values (see Figures 16 and 17 for the analytical and numerical methods, respectively).

Using the deviations from the reference ZWD values, a χ^2_ν was calculated for the two sets of uncertainty measures. For the analytical method, $\chi^2_\nu = 1.5$ was found. The uncertainty measures derived using the numerical method were far too small, resulting in $\chi^2_\nu = 46$.

5. Conclusions

In this paper, we have applied two methods for determining uncertainties in multivariate analysis, numerical and analytical, on two very different test cases.

For the data in Case 1, the analytical and numerical methods produced uncertainties of sizes that are in agreement with the size of the "true errors." For this case, we estimated 28 properties. The χ^2_ν was for most of these between 0.5 and 2.

For the data in Case 2, the analytical method produced uncertainties of sizes that are in agreement with the size of the "true errors." For Case 2, statistics were only evaluated for one property, the ZWD. We found $\chi^2_\nu = 1.5$ using the analytically determined uncertainties. The uncertainty measures derived using the numerical method are underestimated, resulting in $\chi^2_\nu = 46$. This relatively high value is believed to result mainly from large measurement noise compared to mismodelling effects. The improvement of numerical methods for uncertainty determination that better represent different categories of errors is part of future work.

Acknowledgements

This project was partially financed by European Thematic Network "Advanced Mathematical and Computational Tools in Metrology" and by grant 38:10, National Metrology of the Swedish Ministry of Industry, Employment and Communication. Borys Stoew at Onsala Space Observatory, Chalmers University of Technology, provided the microwave radiometer data. The seafood data are part of a project with financial support from the Commission of the European Communities, Fifth Framework Programme, specific RTD programme Quality of Life and Management of Living Resources, project QLK1-2001-01643, "A New Method for Measurement of the Quality of Seafood."

References

1. Bevington, P. R.: *Data Reduction and Error Analysis for the Physical Sciences*, McGraw Hill, New York, 1969.
2. Efron, B. and Gong, G.: A leisurely look at the bootstrap, the jackknife, and the cross-validation, *Am. Stat.* **37** (1983), 36–48.
3. Elgered, G. and Jarlemark, P. O. J.: Ground-based microwave radiometry and long-term observations of atmospheric water vapor, *Radio. Sci.* **33** (1998), 707–717.
4. Faber, N. M.: Uncertainty estimation for multivariate regression coefficients, *Chemometr. Intell. Lab. Syst.* **64** (2002), 169–179.
5. Höskuldsson, A.: PLS regression methods, *J. Chemometr.* **2** (1988), 211–228.
6. Hoy, M., Steen, K. and Martens, H.: Review of partial least squares regression prediction error in unscrambler, *Chemometr. Intell. Lab. Syst.* **44** (1998), 123–133.
7. Johnson, R. A. and Wichern, D. W.: *Applied Multivariate Statistical Analysis*, Pearson Education Int., Upper Saddle River, 2002.
8. Sundberg, R.: Aspects of statistical regression in sensometrics, *Food Qual. Prefer.* **11** (2000), 17–26.
9. Wentzell, P. D., Andrews, D. T. and Kowalski, B. R.: Maximum likelihood multivariate calibration, *Anal. Chem.* **69** (1997), 2299–2311.