Data-driven models for the operation of a dynamic primary standard based on permeation

Federica Gugole[1], *Adriaan M. H.* van der Veen [1,2*], and Ewoud J.J. de Boed[2]

[1]VSL, Data Science & Modelling Department, Thijsseweg 11, 2629 JA Delft, The Netherlands
[2]VSL, Chemistry Department, Thijsseweg 11, 2629 JA Delft, The Netherlands

Abstract. For the preparation of calibration gas mixtures with low levels of reactive compounds, the permeation method as described in ISO 6145-10 is excellently suited. In the past years, it has proved to be useful for developing primary standards for key impurities in hydrogen, such as ammonia, hydrogen fluoride and hydrogen chloride. Using an automated balance to measure the mass loss of the permeation tube, large volumes of data are gathered, from which the permeation mass flow rate is calculated. The residuals of traditional simple straight-line regression show patterns which deserve further attention. They suggest that the current approach may underrate the uncertainty associated with the permeation rate. The permeation system is operated with a dilution system using thermal mass flow controllers. These produce large volumes of data as well, which are processed to obtain a mass flow rate and associated uncertainty. It is shown how time series models enable assessing correlations in the data from the magnetic suspension balance and the thermal mass flow controllers. These data-driven models provide a better description of the features of the data and thereby provide more realistic estimates of the mass flow rates and associated uncertainties. The time series analysis reveals that the permeation data are quite heavily autocorrelated, whereas the data from the thermal mass flow controllers are uncorrelated.

1 Introduction

Permeation is one of the measurement principles used by the gas industry and National Metrology Institutes (NMIs) to create calibration gas mixtures. The method is described in ISO 6145-10 [1] and often applied in conjunction with using thermal mass flow controllers (MFCs) [2] to dose the matrix gas. The method enables preparing calibration gas mixtures with the component of interest at µmolmol^{-1}-level in one step. The method can be used for a wide range of components including sulfur dioxide, nitrogen dioxide [3], hydrogen chloride, hydrogen fluoride and formaldehyde [4]. The method was successfully applied in key comparisons [5, 6] and for developing a primary measurement standard for key impurities in hydrogen [4, 7], among many other applications.

Especially the use of a magnetic suspension balance (MSB) to continuously weigh the permeation tube enables high-end applications. The MSB is used to record the mass loss as a function of time. The data are used as the mass flow rate of the component of interest. The permeation tube is filled with a gas, liquid or solid of adequate purity that permeates through the wall. The substance that permeates through the wall is diluted with the matrix gas(es) and carried to the analyser. A typical data set from the permeation of hydrogen chloride is shown in figure 1. The slope of the fitted line is the permeation rate of the hydrogen chloride.

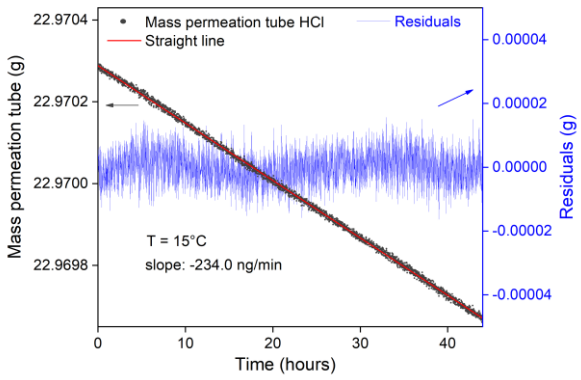

Fig. 1. Permeation data of hydrogen chloride at 15 °C. The results (fitted straight line and residuals) of OLS regression are also displayed.

The patterns in the residuals from fitting a straight line through the permeation data reveal that they are not independent and identically distributed (IID). These patterns hint at the presence of autocorrelation in the data. Autocorrelation means that a datum in a time series depends on its predecessor or a few predecessors [8, 9]. Just as with other kinds of dependencies, the presence of autocorrelation in a data set influences any results calculated from it.

* Corresponding author: avdveen@vsl.nl

An MSB generates a lot of data, equally spaced in time, which lends itself well for the application of time series analysis. In this work, time series analysis is applied to (1) assess the structure of the data and (2) specify a model that describes the dependencies in the data. Then, the time series analysis is combined with the regression to replace the ordinary least squares (OLS) commonly used when obtaining the permeation rate.

2 Analysis

2.1 Time series

Differently from a more general data set, in a time series the (temporal) order in which the data have been recorded matters. Time series models thus appreciate the temporal relation between subsequent data points and enable modelling phenomena with a temporal behaviour and capture any resulting correlations. More classical statistical techniques such as OLS instead, are not sensitive to the order in which the data are stored in the data set, and therefore they cannot capture any temporal dependency among the data.

Measurement data collected using a dynamic method, such as permeation, naturally have a temporal component and the order in which these data are collected matters since it gives information on the behaviour of the system. Therefore, data sets collected by such systems are natural candidates for time series analysis.

2.2 MSB data set

The first data set analysed in this work is the data set from the permeation of hydrogen chloride shown in figure 1. The data set covers more than 40 hours of measurements recorded on average every 0.67 min. The residuals of the OLS regression displayed in figure 1 (blue line) show an oscillating behaviour which hints that they are not IID. Factors that may contribute in this regard are the temperature control and the pressure regulator. The permeation chamber is temperature and pressure controlled, thus the heater switches on and off when the temperature reaches predefined thresholds, and a similar behaviour is expected also by the pressure regulator.

The MSB performs an automatic zeroing after every sixth measurement. Therefore, the time interval between the sixth and seventh measurement is about 2.23 min. Since time series analysis requires the data to be equally spaced in time, the question arises whether these gaps should be filled in the series to have a more homogeneous temporal spacing over the whole series. However, this would require imputing two data points in each auto-zeroing interval. Consequently, imputing would lead to 25 % of the final time series to be imputed data with statistical properties potentially different with respect to the original data. Given the large percentage of imputed data, the results of the analysis might be significantly affected by the chosen imputing method. Therefore, it was decided to analyse the dataset as is and to not impute any data.

2.3 MFC data set

The second data set considered here is a data set collected during a stability test of the MFCs used in the permeation standard. There, an MFC with 100 mL min^{-1} capacity was set to a volume flow rate of 50 mL min^{-1}. The actual volume flow rate was recorded from the MFC sensor and from a piston prover. The test lasted about twenty hours and data were recorded every movement of the piston and then averaged per 10 measurements. The data set containing the MFC readings is shown in figure 2.

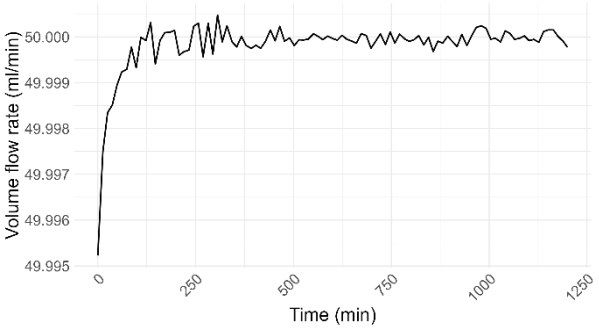

Fig. 2. Observed volume flow rate recorded by a mass flow controller during a stability test.

After an initial warm-up of the MFC, the recorded flow rate oscillates around a constant value. In the time series analysis the initial warm-up phase is neglected as it is clear that in that phase the properties of the system are different from the remaining series of data, thus breaking the assumption of (weak) stationarity, see [8, 9] for more information. The warm-up and stabilization were confirmed by the measurements using the calibrated piston prover.

2.4 Time series model

When fitting the MSB dataset with OLS, the so-obtained residuals show some visible structure, see figure 1. This indicates that the assumptions for the OLS algorithm are not satisfied and that other statistical techniques should be used to perform the regression. In particular, if the assumption that the error terms are uncorrelated is broken, the OLS algorithm will tend to underestimate the standard errors associated to the parameters.

In this work, the so-called seasonal autoregressive integrated moving average (SARIMA) model [8, 9], a generalisation of the autoregressive moving average (ARMA) model to periodic variations, was applied to the residuals of the OLS regression following a procedure similar to the one described in [10, 11] for ARMA processes. SARIMA models allow to model both the seasonal and the within season behaviour of the data as an ARMA model, where the order of each model is to be determined as described in [10] and keeping in mind that only the autocorrelation function (ACF) and partial ACF (PACF) entries at multiples of the seasonality are to be considered when determining the model for the seasonal part. Lastly, a straight-line with autocorrelated residuals was fitted to the data and the result compared with the outcome of the OLS regression.

The MFC dataset does not display visible structure. Therefore, in this case time series analysis is used to check if there is any hidden pattern causing (auto-) correlation or if the assumption of IID values holds.

3 Results

3.1 Regression with autocorrelated residuals

Both the ACF and PACF of the OLS residuals show significant correlation every sixth data point, see figure 3. This coincides with the timings of the automatic zero autocalibration regularly performed by the MSB. Thus, it is reasonable to model this as a seasonal (or cyclic) event with a given seasonality length of six.

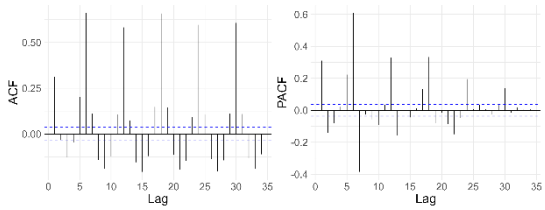

Fig. 3. ACF and PACF values of the residuals obtained fitting a straight line to the MSB data set using OLS.

Since the correlation peaks due to the seasonality are more or less all of the same value in the ACF but are decreasing in the PACF, it was decided to model the seasonality as an autoregressive (AR) model of order 1. This is not the only correlation present in the data though. Within the six-measurement cycle it is possible to note that measurements at distance one, i.e., measurements that are recorded subsequently after each other, are also correlated. The PACF displays less significant correlations within a single season, so an AR model of order 1 is used also for the within season model component.

The residuals of the resulting SARIMA model do not show any visible structure and their ACF and PACF show much less correlation, see figure 5. In particular, the peaks at regular intervals have disappeared, indicating the seasonal component has been correctly captured by the model. The oscillating pattern visible in the OLS residuals has also disappeared (see figure 4) thus it has also been captured by the SARIMA model. Both the ACF and PACF still show some significant correlations with higher terms (i.e., data points with higher temporal separation between each other). However, no clear pattern is visible and the correlation values are rather low (approximately 0.2 in absolute value), so it is concluded that this correlation is mostly due to noise and it is not necessary to further complicate the model.

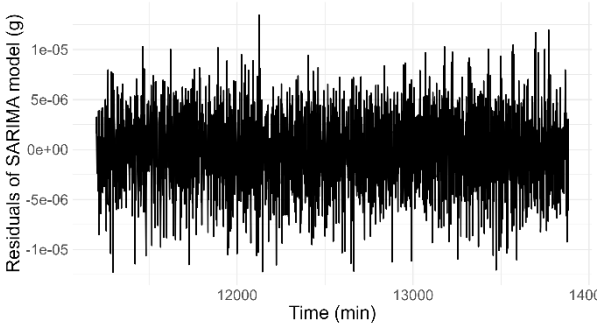

Fig. 4. Plot of the residuals of the SARIMA model applied to the OLS residuals.

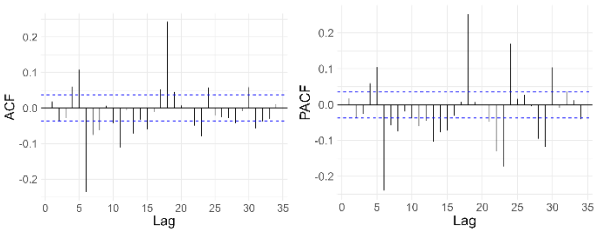

Fig. 5. Plot of the ACF and PACF of the residuals of the SARIMA model applied to the OLS residuals.

The SARIMA model for the residuals can thus be included in the regression algorithm. In table 1 the regression parameters and associated standard errors calculated with or without correlation among the OLS residuals are reported. The mean estimate of the parameter values does not change significantly but the standard error is almost four times larger when including the SARIMA model for the residuals.

Table 1. Regression coefficients value and associated standard error obtained either using OLS (i.e., assuming independence of the residuals) or including a SARIMA model for the residuals.

	Coefficient	Value	Std. error
OLS	Intercept	22.97	$1.64 \cdot 10^{-6}$
	Slope	$-2.340 \cdot 10^{-7}$	$1.31 \cdot 10^{-10}$
SARIMA	Intercept	22.97	$6.06 \cdot 10^{-6}$
	Slope	$-2.340 \cdot 10^{-7}$	$5.15 \cdot 10^{-10}$

3.2 Stability of MFC

Since no visible structure in the MFC data set is visible in figure 2, the ACF and PACF values (see figure 6) may be checked to ensure that the IID assumption can be used for this case.

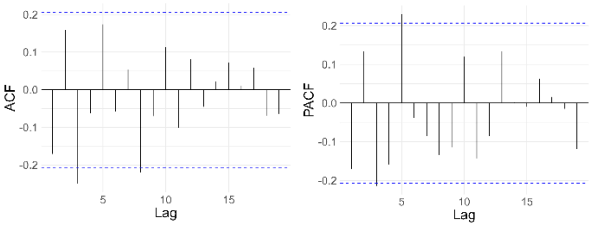

Fig. 6. ACF and PACF values of the MFC data recorded during a stability test.

Most of the ACF and PACF are not statistically different from zero since they are within the 95 % confidence interval defined by the blue lines. Those

whose value exceeds the limit of the confidence interval are only barely outside, with a correlation value rather low (about 0.2) and they are at higher terms. It was thus concluded that there is indeed no underlying structure in the data, and the data can be treated as independent.

4 Conclusions

Dynamic methods for preparing calibration gas mixtures, such as permeation, generate data sets where the order in which the data have been recorded matters. Time series models thus are natural candidates for the analysis of such data sets since they can capture the temporal relationship between the measurement results and thus determine any potential correlation.

This work showed two practical examples of how time series analysis can effectively be used to estimate (auto-) correlations, if any, between measurements. In particular, their application to a data set from the permeation of hydrogen chloride recorded by a magnetic suspension balance allowed to incorporate the regular automatic zero-recalibration performed by the balance in the regression model estimating the permeation rate. This resulted in an almost four times larger standard error associated to the parameters estimated by the regression model with respect to the case where simple straight-line regression is used.

The application of the (partial) autocorrelation function to a data set collected by a mass flow controller during a stability test revealed that for such a test the assumption of IID measurement results is acceptable and may thus be used in the uncertainty analysis.

In both cases, the results of the time series analysis can be readily included in the uncertainty budget of the dynamic method. In the case of the analysis on the MSB data sets, the standard error associated with the permeation rate estimated by the time series model would have be used instead of the one obtained used OLS. In the case of the analysis on the data set recorded by the MFCs, no changes are required in the uncertainty budget since in that case the time series analysis simply confirmed that the assumption of IID values holds.

Time series analysis is therefore a useful tool in the analysis of data sets recorded by dynamic methods or systems. More work on how to combine and propagate measurement uncertainty in the presence of serial correlation is under way in the project 24GRD10 SmartGasNet.

Acknowledgment

This project has received funding from the Ministry of Economic Affairs of the Netherlands.

Statements

The data that support the findings of this study are available from the corresponding author upon request. The data are provided exclusively for non-commercial research purposes.

Author contributions
Conceptualization: FG, AvdV; Methodology: FG; Software: FG; Validation: FG; Formal analysis: FG; Investigation: FG; Resources: AvdV, EdB; Writing- Original draft: FG, AvdV, EdB; Visualization: FG, AvdV; Supervision: AvdV; Project administration: AvdV; Funding acquisition: AvdV

References

1. ISO 6145-10 Gas analysis – Preparation of calibration gas mixtures using dynamic volumetric methods – Part 10: Permeation methods (2002).
2. ISO 6145-7 Gas analysis – Preparation of calibration gas mixtures using dynamic volumetric methods – Part 7: Thermal mass-flow controllers (2018).
3. E. Flores, J. Viallon, P. Moussay, F. Idrees, and R. I. Wielgosz. Highly accurate nitrogen dioxide (NO2) in nitrogen standards based on permeation. Anal. Chem., **84**(23):10283 (2012).
4. H. Meuzelaar, J. Liu, S. Persijn, J. van Wijk, and A. M. H. van der Veen. Trace level analysis of reactive ISO 14687 impurities in hydrogen fuel using laser-based spectroscopic detection methods. Int. J. Hydrogen Energy, **45**(58):34024-34036 (2020).
5. E. Flores, et al.. International comparison CCQM-K74.2018: Nitrogen dioxide, 10 µmol/mol. Metrologia, **58**(1A):08018 (2021).
6. J. Viallon, E. Flores, F. Idrees, P. Moussay, R. I. Wielgosz, D. Kim, Y. D. Kim, S. Lee, S. Persijn, L. A. Konopelko, Y. A. Kustikov, A. V. Malginov, I. K. Chubchenko, A. Y. Klimov, O. V. Efremova, Z. Zhou, A. Possolo, T. Shimosaka, P. Brewer, and T. Macé. CCQM-K90, Formaldehyde in nitrogen, 2µmol mol^{-1} Final report. Metrologia, **54**(1A):08029-08029 (2017).
7. ISO 14687 Hydrogen fuel quality – Product specification (2025).
8. C. Chatfield and H. Xing. The analysis of time series. Chapman & Hall / CRC Texts in Statistical Science. CRC Press, London, England, 7th edition (2019).
9. R.H. Shumway and D.S. Stoffer. Time series analysis and its applications: with R examples. EDP Springer International Publishing, (2017).
10. A. M. H. van der Veen, F. Gugole, K. Folgerø, A. M. Skålvik, J. Kutin, G. Bobovnik, K. Rasmussen, L. C. Nordhjort Mjølna, and E. Venslovas. Best practices in the evaluation of the measurement uncertainty of quantities relevant to fiscal measurements along the hydrogen supply chain. (2025).
11. A. M. H. van der Veen, F. Gugole, K. Folgerø, A. M. Skålvik, J. Kutin, G. Bobovnik, K. Rasmussen, L. C. Nordhjort Mjølna, and E. Venslovas. Metering uncertainty for custody transfer of hydrogen for transport, heat, and storage. (2025).