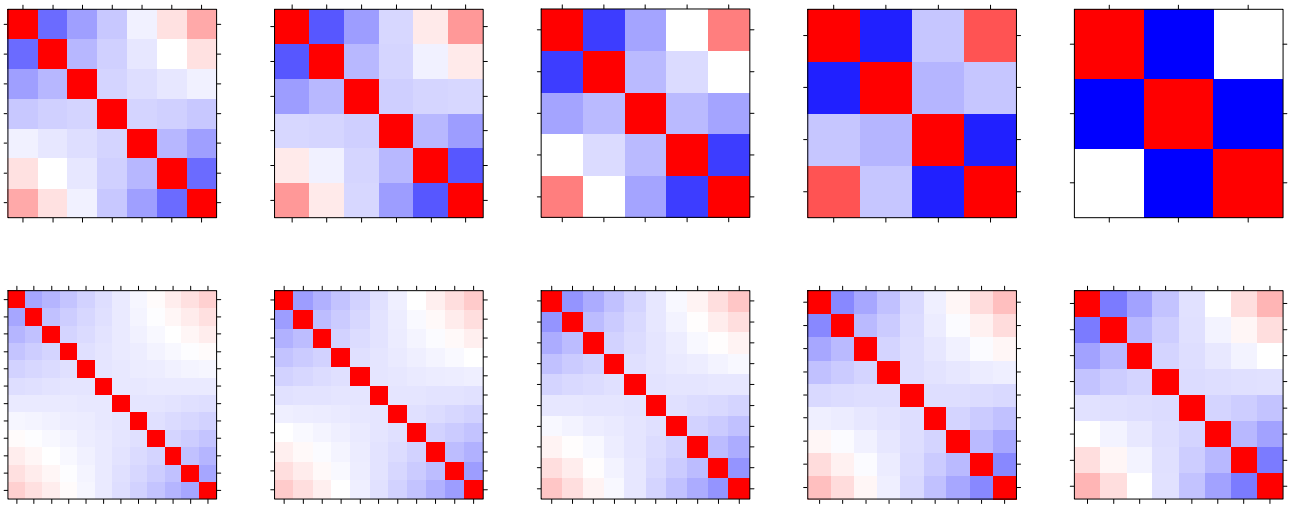
Environment
Agroscope Science | No. 43 / 2017



Testing Monitoring Data for Serial Correlation

Author
Andreas Gubler

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
Agroscope

Impressum

Editor:	Agroscope Reckenholzstrasse 191 8046 Zürich www.agroscope.ch
Information:	Andreas Gubler, andreas.gubler@agroscope.admin.ch
Cover	Correlation matrices for repeated measures
Copyright:	© 2017 Agroscope
Download:	www.agroscope.ch/science
ISSN:	2296-729X
ISBN:	978-3-906804-31-6

Table of contents

List of figures

Abstract

Soil monitoring programs, such as the Swiss Soil Monitoring Network NABO, typically generate datasets consisting of data points separated by few years measured repeatedly at the same ensemble of monitoring sites. For the statistical analyses, it is important to know if the errors are correlated in time. However, even in the case of independent errors (that are unknown) the model residuals are correlated. This report presents a method using residual analysis to test the null hypothesis of independent errors. The method is illustrated for data on the temporal evolution 1990–2014 of organic carbon contents for 29 cropland sites. The results indicated no temporal correlation for the assessed data.

Zusammenfassung

Bodenmonitoringprogramme, wie die Nationale Bodenbeobachtung NABO, generieren üblicherweise Datensätze mit mehreren Datenpunkten pro Standort aus wiederholten Probenahmen, die einige Jahre auseinanderliegen. Für die statistischen Auswertungen ist es von Bedeutung, ob die Fehler der verschiedenen Zeitpunkte korreliert sind. Allerdings sind die Modellresiduen selbst im Falle unkorrelierter Fehler (die per se unbekannt sind) korreliert. Dieser Bericht zeigt wie mit Hilfe von Residuenanalysen die Hypothese, die Fehler seien unabhängig, überprüft werden kann. Die Methode wird an Daten zur zeitlichen Entwicklung 1990–2014 von Kohlenstoffgehalten auf 29 Ackerstandorten demonstriert. Für die untersuchten Daten fanden wir keine Hinweise auf zeitliche Korrelationen.

Résumé

Les programmes de monitoring du sol comme l'Observatoire national des sols NABO génèrent généralement des séries de données comprenant plusieurs points de données prélevés sur le même site et à plusieurs années d'intervalle. Pour les analyses statistiques, il est important de savoir si les erreurs sont corrélées dans le temps, Toutefois, même dans le cas d'erreurs indépendantes (qui ne sont pas connues en soi), les résidus du modèle statistique sont corrélés. Ce rapport présente une méthode permettant de tester à l'aide d'analyses résiduelles l'hypothèse selon laquelle les erreurs sont indépendantes. La méthode est illustrée par les données de l'évolution des teneurs en carbone dans le temps (de 1990 à 2014) sur 29 sites cultivés. Les résultats n'indiquent aucune corrélation temporelle pour les données étudiées.

1 Introduction

Data sets generated by environmental monitoring programs strongly differ regarding their temporal resolution. Whereas some record their target variables (almost) continuously, e.g. air quality programmes, many others generate data points separated by years. The latter setting is typical for soil monitoring programmes, such as the Swiss Soil Monitoring Network (NABO; Gubler et al. 2015). NABO operates 110 long-term monitoring sites throughout Switzerland covering all relevant land uses. Most sites were sampled for the first time between 1985 and 1989 and re-sampled every five years ever since. For each sampling campaign and site, four replicate samples are collected from the top 20 cm of soil for chemical analyses. Usually, statistical analyses are based on the mean of the four replicates. Hence, the analysed datasets comprise observations from a specific number of sites, each measured at several (typically five to six) almost equidistant time points. Such data may be analysed as repeated measures, e.g. by using linear-mixed models or other hierarchical approaches. If the errors of the model exhibit temporal correlation (such as serial correlation), we have to account for it.
The errors are the differences between the true values and the fitted ones. However, the true values (and hence the errors) are unknown since our data represent only estimates of the true values distorted by various error sources (sampling errors, analytical uncertainty …). Therefore, one can only inspect the model residuals that differ from the errors with respect to some properties. In particular, even in the case of independent errors, residuals are correlated. For simplicity's sake, we will use a simple linear regression for example. If there is one predictor x_t for the response variable y_t determined for *N* time points, the model equation is

$$y_t = \beta_0 + \beta_1 \cdot x_t + \varepsilon_t \qquad t = 1, \ldots, N \tag{1}$$

or, using matrices,

$$y = X\,\beta + \varepsilon \qquad X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{2}$$

where $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are column vectors containing the response and the errors, respectively. $\boldsymbol{X}$ represents the so-called design matrix. The regression coefficients $\boldsymbol{\beta}$ are estimated from the data and used to derive the fitted values:

$$\hat{y} = X\,\hat{\beta} \tag{3}$$

Then, the residuals are the differences between our observations and the fitted values:

$$\hat{\varepsilon} = y - \hat{y} \tag{4}$$

In the case of independent errors, their covariance is

$$Cov(\varepsilon) = I\,\sigma^2 \tag{5}$$

where $\boldsymbol{I}$ is a *N* x *N* identity matrix. In contrast, the covariance of the residuals is

$$Cov(\hat{\varepsilon}) = (I - H)\sigma^2 \qquad H = X(X^TX)^{-1}X^T \tag{6}$$

The residuals are correlated because $\boldsymbol{H}$ is not a diagonal matrix (and thus the non-diagonal elements of the covariance matrix differ from zero). The theoretical correlations of a linear model according Equation 1 including five time points are displayed in Figure 1 (see cover graphic for correlations of datasets featuring less/more time points). Most remarkable are the negative correlations of the errors of the first vs. the second and the fourth vs. the fifth time point. One can demonstrate this phenomenon by drawing randomly five points and then adding a trend line. It is very likely that the line passes between the first and the second point producing a positive residual for the first and a negative residuals for the second point, or vice versa. The same holds for the fourth and the fifth point. Hence, negative correlations are likely to be observed for the first and the second respectively the fourth and fifth point.

In summary, we know in advance that the residuals are correlated, even when the errors are not. Thus, we must assess if the residuals are correlated as we would expect under the null hypothesis of independent errors. This report presents a method using residual analysis to test whether errors are correlated or not. Data on organic carbon (OC) contents determined for NABO monitoring sites are used for illustration. The used methodology as well as the theoretical background outlined above was elaborated by Stahel et al. (2014) in the context of previous analyses of monitoring data collected by NABO.

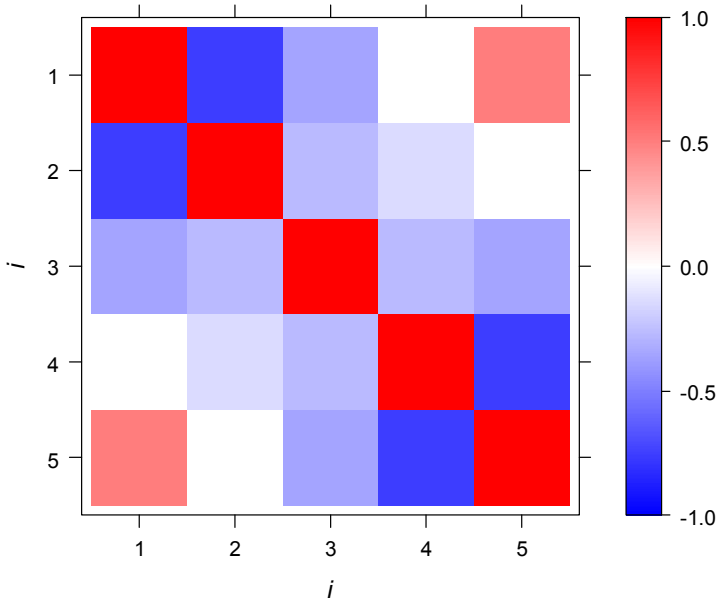


Figure 1: Theoretical correlation matrix for a first order linear model including five time points.

2 Methods

2.1 Data

We used data on OC contents measured repeatedly at 29 NABO monitoring sites used as cropland (Gubler et al. 2017). All sites were sampled five times between 1990 and 2014 (NABO sampling campaigns 2 to 6; results of campaign 1 were omitted since we assumed that they are biased). Samplings at the individual sites were separated by five years. The statistical analyses were based on the mean of four replicates per site and sampling, OC contents were log-transformed.

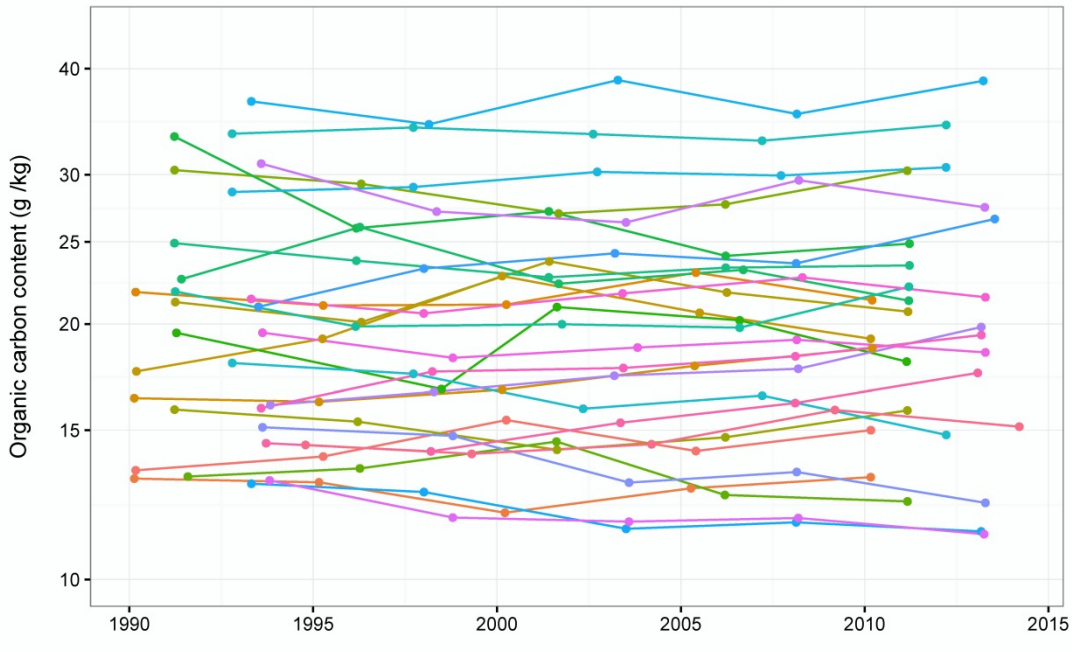


Figure 2: Data on organic carbon contents of top 20 cm of 29 cropland sites sampled repeatedly five times from 1990 to 2014 by the Swiss Soil Monitoring Network NABO.

2.2 Residual analyses

The used method included the steps below. Correlation matrices were calculated by using Spearman's rank correlation method to reduce the impact of outliers.

0. Remove sites with missing observations (the dataset described above only included monitoring sites with complete observations).
1. Fit a linear regression for every site. The linear model can be written as $y_{i,t} = \beta_{0,i} + \beta_{1,i} \cdot t + \varepsilon_{i,t}$ where $\boldsymbol{y_{i,t}}$ represents the log of the OC content at site *i* for time point *t*, $\beta_{0,i}$ represent the site-specific OC levels, and $\beta_{1,i}$ allows for an individual slope per site.
2. Calculate the residuals of the regression ("observed residuals") and derive the corresponding correlation matrix between the different time points ("observed correlation matrix"; in our case a 5 x 5 matrix).
3. Generate independent residuals by simulating from a *t*-distribution with 3 degrees of freedom (which was chosen based on empirical inspection of data, i.e. by using quantile-quantile plots; alternatively, one could simply permutate the residuals obtained in step 2). Fit the model used in step 1 to the simulated data, calculate the residuals ("simulated residuals"), and derive the correlation matrix ("simulated correlation matrix").
4. Iterate the previous step 1000 times. This ensemble of simulated correlation matrices represent the correlations we may expect under the null hypothesis of independent errors.
5. Calculate the theoretical correlation matrix using the design matrix (c.f. Equation 6).
6. Compute a *p*-value for each cell of the correlation matrix as follows: Calculate the absolute deviation of the observed correlation from the theoretical correlation and compare with the absolute deviations of the simulated correlations. The proportion of simulations with absolute deviations larger than the one of the real data represents the *p*-value.
7. Compute an overall *p*-value by using only the correlations at lag 1 (time point 1 vs. 2, 2 vs. 3, etc.). The corresponding correlations are squared and summed to derive the test statistic. The test statistic of the real data is compared to those of the simulations. The proportion of simulations with test statistics larger than the one of the real data represents the *p*-value. (Hence, at a significance level of 5 %, we reject the null hypothesis of independent errors if the test statistic of the real data is more extreme than 95 % of the simulations).

3 Results, Discussion & Conclusion

The scatter plots of the observed residuals (i.e. the residuals of the real data) indicated negative correlations for sampling campaign 2 vs. 3 and campaign 5 vs. 6 (Figure 3).These findings were confirmed by the observed correlation matrix (Figure 4), although the correlations were less negative than expected from the theoretical correlation matrix (Figure 1). The difference of the observed and the theoretical correlations was most positive for the correlation of campaign 2 vs. 3 (Figure 5). In contrast, some correlations at lag 2 were more negative than expected. However, the observed deviations were in good agreement with those of the simulated data. The *p*-values of the individual cells of the correlation matrix were considerably larger than the applied significance level of 0.05 (Figure 6). The lowest *p*-value of 0.21 was observed for the correlation of campaign 2 vs. 3. The overall test statistic of the real data was similar to the test statistics of the simulated data, the corresponding *p*-value was 0.81 (Figure 7). Accordingly, the null hypothesis of independent errors was not rejected.

In conclusion, the errors of the different sampling campaigns seem not to be correlated for the assessed OC data. Temporal correlation might be relevant for shorter time lags, in particular for yearly or even more frequent samplings. The presented method may be used (and potentially adapted) for similar datasets.

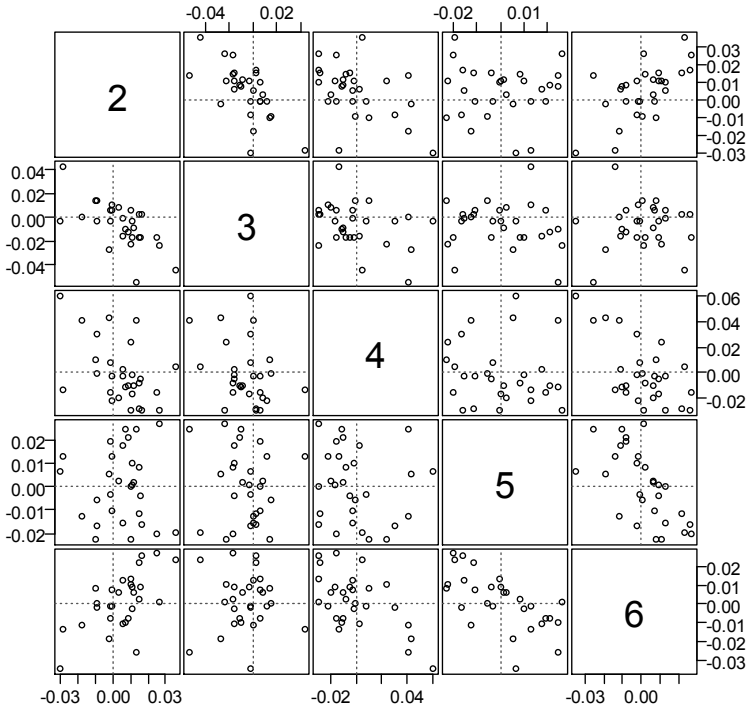


Figure 3: Model residuals observed for OC data.

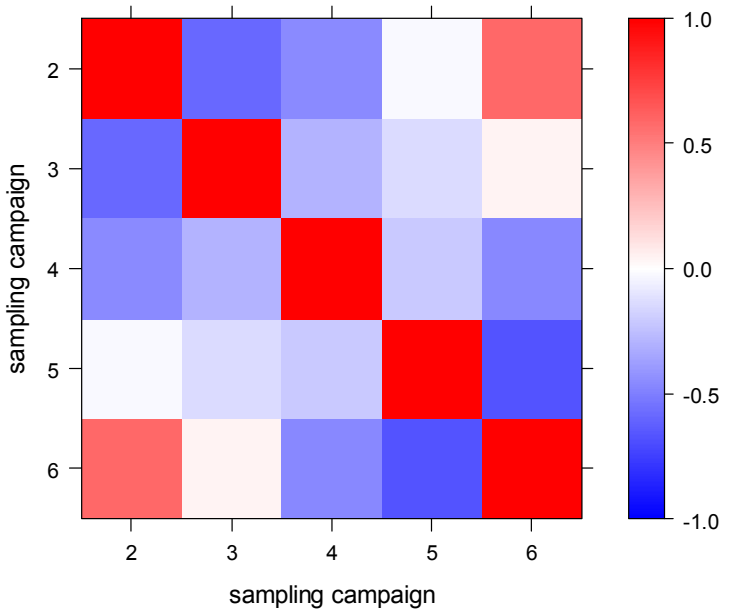


Figure 4: Correlation matrix observed for residuals of OC data.

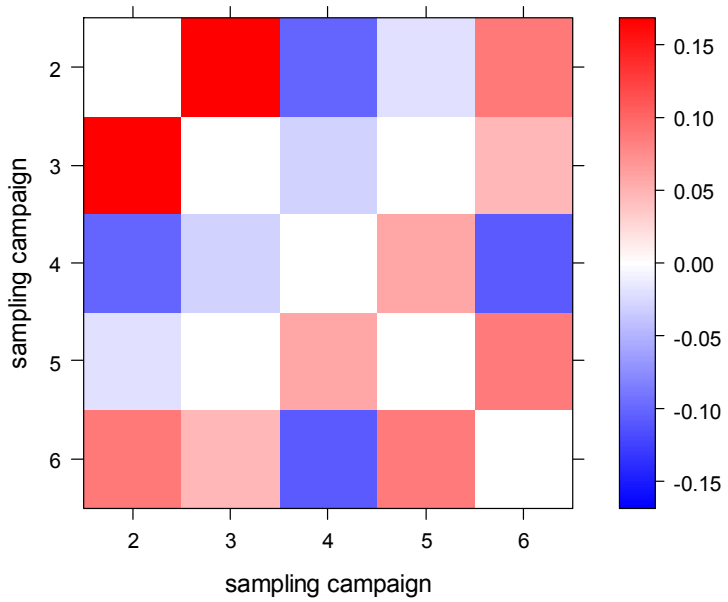


Figure 5: Deviations of the observed correlations (Figure 4) from the theoretical correlations (Figure 1).

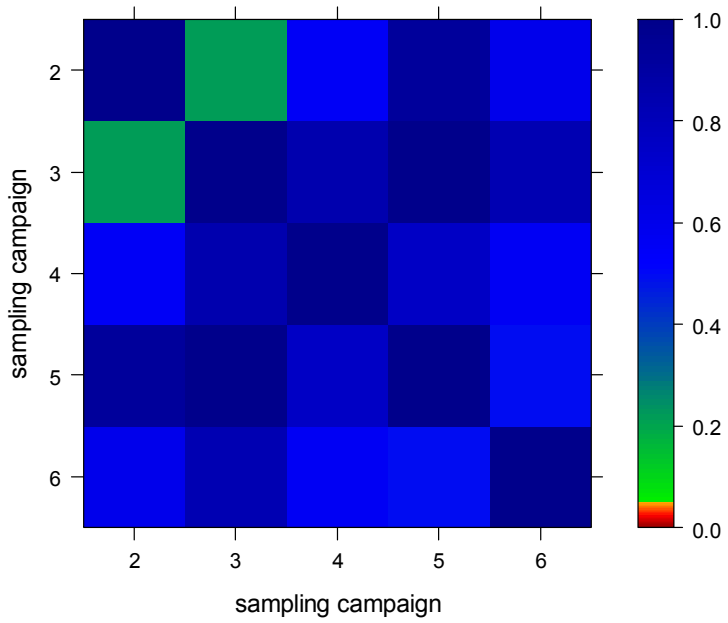


Figure 6: p-values of the individual cells of the correlation matrix.

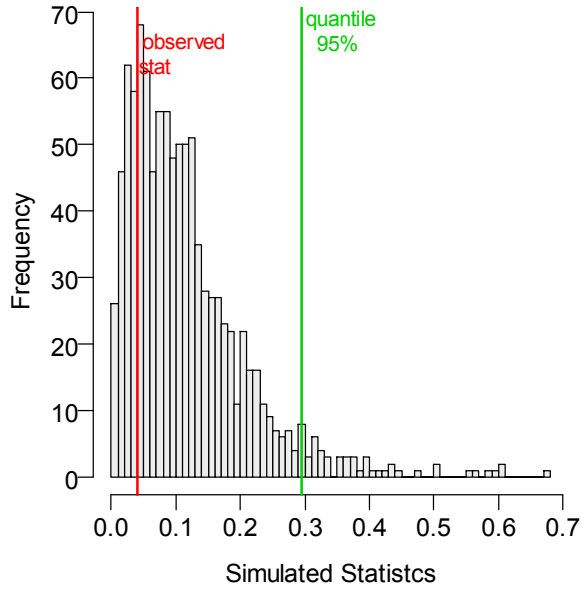


Figure 7: Overall test statistic based on lag 1 correlations of the real data (red line) compared with test statistics of the simulations (histogram) and their 0.95 quantile (green line).

4 References

Gubler A, Schwab P, Wächter D, Meuli R G, Keller A (2015) Ergebnisse der Nationalen Bodenbeobachtung (NABO) 1985-2009. Federal Office for the Environment (FOEN). Umwelt-Zustand Nr. 1507. Bern.

Gubler A, Schwab P, Wächter D, Müller M, Keller A (2017; *in preparation*). Organic carbon contents of mineral cropland soils in Switzerland remained stable over the last 30 years. *In preparation*.

Stahel W, Tanadini M, Hannay M (2014). Bodenbeobachtung Analyse. Internal report on behalf of Swiss Soil Monitoring Network NABO, Seminar for Statistics, ETH Zurich.

Acknowledgements

The NABO is co-financed by the Federal Office for the Environment (FOEN) and Federal Office for Agriculture (FOAG). For the present study, we particularly acknowledge the efforts by Werner Stahel, Matteo Tanadini, and Mark Hannay, Seminar for Statistics, ETH Zurich.