Journal of Hydrology 540 (2016) 527–537

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Detecting spatial structures in throughfall data: The effect of extent, sample size, sampling design, and variogram estimation method

Sebastian Voss [a], Beate Zimmermann [b], Alexander Zimmermann [a,*]

[a] University of Potsdam, Institute of Earth and Environmental Science, Potsdam, Germany
[b] Research Institute for Post-Mining Landscapes, Finsterwalde, Germany

ARTICLE INFO

Article history:
Received 5 October 2015
Received in revised form 10 May 2016
Accepted 19 June 2016
Available online 21 June 2016
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Uwe Haberlandt, Associate Editor

Keywords:
Throughfall
Geostatistics
Sampling
Variogram
Residual maximum likelihood

ABSTRACT

In the last decades, an increasing number of studies analyzed spatial patterns in throughfall by means of variograms. The estimation of the variogram from sample data requires an appropriate sampling scheme: most importantly, a large sample and a layout of sampling locations that often has to serve both variogram estimation and geostatistical prediction. While some recommendations on these aspects exist, they focus on Gaussian data and high ratios of the variogram range to the extent of the study area. However, many hydrological data, and throughfall data in particular, do not follow a Gaussian distribution. In this study, we examined the effect of extent, sample size, sampling design, and calculation method on variogram estimation of throughfall data. For our investigation, we first generated non-Gaussian random fields based on throughfall data with large outliers. Subsequently, we sampled the fields with three extents (plots with edge lengths of 25 m, 50 m, and 100 m), four common sampling designs (two grid-based layouts, transect and random sampling) and five sample sizes (50, 100, 150, 200, 400). We then estimated the variogram parameters by method-of-moments (non-robust and robust estimators) and residual maximum likelihood. Our key findings are threefold. First, the choice of the extent has a substantial influence on the estimation of the variogram. A comparatively small ratio of the extent to the correlation length is beneficial for variogram estimation. Second, a combination of a minimum sample size of 150, a design that ensures the sampling of small distances and variogram estimation by residual maximum likelihood offers a good compromise between accuracy and efficiency. Third, studies relying on method-of-moments based variogram estimation may have to employ at least 200 sampling points for reliable variogram estimates. These suggested sample sizes exceed the number recommended by studies dealing with Gaussian data by up to 100 %. Given that most previous throughfall studies relied on method-of-moments variogram estimation and sample sizes ≪200, currently available data are prone to large uncertainties.

1. Introduction

In the last three decades, an increasing number of studies analyzed spatial patterns in throughfall to investigate the consequences of rainfall redistribution for biogeochemical (Allen et al., 2015; Hsueh et al., 2016; Möttönen et al., 1999; Whelan et al., 1998) and hydrological processes in forests (Fathizadeh et al., 2014; Gerrits et al., 2010; Hsueh et al., 2016; Keim et al., 2005; Klos et al., 2014; Loescher et al., 2002; Shachnovich et al., 2008; Staelens et al., 2006; Zimmermann et al., 2009). Other studies analyzed throughfall spatial patterns to optimize sampling schemes for estimating mean throughfall (Ziegler et al., 2009; Zimmermann and Zimmermann, 2014; Zimmermann et al., 2010). In the majority of cases, variograms were used to characterize the spatial properties of the throughfall data.

The variogram is central to geostatistics (Webster and Oliver, 1992, 2007) because it describes spatial variation and provides the parameters (nugget, sill and range) that are essential for spatial prediction and the simulation of random fields. It is widely accepted that estimates of the variogram are sensitive to the size and spatial arrangement of the sample (e.g. Lark, 2002a; Russo and Jury, 1987; Webster and Oliver, 1992). Furthermore, there is ample evidence that the variogram range and sill depend on the spatial scale of sampling (e.g. Blöschl, 1999; Western and Blöschl, 1999). Several studies investigated the influence of various aspects of the sampling design on variogram estimation (Bogaert and Russo, 1999; Blöschl, 1999; Corsten and Stein, 1994; Kerry et al.,

* Corresponding author.
E-mail address: alexander.zimmermann.ii@uni-potsdam.de (A. Zimmermann).

http://dx.doi.org/10.1016/j.jhydrol.2016.06.042
0022-1694/

2008; Lark, 2002a; Morris, 1991; Müller and Zimmerman, 1999; Pardo-Igúzquiza and Dowd, 2013; Pettitt and McBratney, 1993; Russo and Jury, 1987; Skøien and Blöschl, 2006; Warrick and Myers, 1987; Webster and Oliver, 1992; Western and Blöschl, 1999). This work, however, has received little attention among forest hydrologists partly because of a missing common language between environmental statisticians and field hydrologists.

A closer look at the studies that investigated the role of sampling designs on variogram estimation reveals that they can be divided into three groups. The first group (Bogaert and Russo, 1999; Lark, 2002a; Morris, 1991; Müller and Zimmerman, 1999; Pettitt and McBratney, 1993; Warrick and Myers, 1987) optimized sampling designs based on various criteria linked to the variogram. Early studies (Morris, 1991; Warrick and Myers, 1987) focused on the distribution of sampling points among the lags. For instance, Warrick and Myers (1987) presented a criterion that aims on an equally distributed number of paired comparisons in each lag class. Morris (1991) criticized this approach because it neglects the correlation of the spatial data and leads to a comparatively low efficiency of the sampling design (van Groenigen, 1999). Subsequent studies chose other, more complex criteria. For instance, Müller and Zimmerman (1999) maximized the determinant of Fisher's information matrix and Lark (2002a) minimized the kriging variance to find an optimum configuration of sampling points. In his comprehensive study, Lark (2002a) demonstrated that (i) a random process with a small spatial dependence is sampled best with scattered clusters of sampling points, (ii) for long range processes a regular array is optimal, and (iii) if there is no prior information about the spatial correlation, sampling in transects is the most robust approach.

The second group of studies (Corsten and Stein, 1994; Kerry et al., 2008; Pardo-Igúzquiza and Dowd, 2013; Webster and Oliver, 1992) sampled simulated random fields to assess the effect of different sampling designs and sample sizes on variogram estimation. This simple approach has the advantage that the experimental variogram can be compared directly against the variogram which is based on all data of the simulated field. Webster and Oliver (1992) sampled different random fields and concluded that a sample size of 150 would be satisfactory for a precise estimate of the variogram. Kerry et al. (2008) followed the approach of Webster and Oliver (1992) and compared residual maximum likelihood (REML) with method of moment (MoM) based variogram estimation. They found that REML outperforms MoM and that a sample size of 100 would be sufficient for variogram parameter estimation.

The third group of studies (Blöschl, 1999; Skøien and Blöschl, 2006; Western and Blöschl, 1999) investigated effects of scale on variogram estimation. Although these studies did not primarily focus on the influence of the sampling design on variogram estimation, their work has important implications for sampling. For instance, Western and Blöschl (1999) showed that estimates of the correlation length depend on the extent (i.e. on the size of the sampling plot).

Most of the studies that assessed the impact of the sampling design on variogram estimation worked with normally distributed data. Furthermore, the data of previous studies often showed a comparatively strong autocorrelation and a long range. Throughfall data, however, usually do not follow a normal distribution; instead they often show skewed underlying distributions (Zimmermann and Zimmermann, 2014) and heavily outlying values (Lloyd and Marques, 1988; Zimmermann et al., 2009). Moreover, variograms of throughfall data usually exhibit relatively small ranges compared to the extent of the research area (Möttönen et al., 1999; Zimmermann and Zimmermann, 2014; Zimmermann et al., 2009). Therefore, it is not clear if the results of previous studies apply for the spatial analysis of data with these properties. To fill this knowledge gap, we sampled a set of unconditional simulations, which we obtained using real-world throughfall data, with several extents, common spatial sampling designs, and a variety of sample sizes. We then evaluated these sampling schemes in terms of their ability to provide satisfactory estimates of the variogram parameters. For our analysis we used both REML and MoM variogram estimation.

2. Methods

2.1. Data

From a large throughfall data set (Zimmermann and Zimmermann, 2014), we selected six events that showed distinct univariate distributions and autocorrelation structures, respectively. While all events included outlying values (i.e. data points which cannot be forced to the center of the distribution even after transformation), events 1 and 5 furthermore showed an underlying asymmetry (cf. Kerry and Oliver, 2007). This type of asymmetry is not caused by outliers but by a skew of the underlying (or primary) distribution of the data, which can be statistically defined as the region between the first and seventh octile (Zimmermann et al., 2010). It is important to distinguish between underlying asymmetry and skewness due to outliers because these deviations from the normal distribution require different treatments of the data. Robust variogram estimators can deal with normally distributed data that are contaminated with outliers. Robust estimators, however, cannot deal with data that show an underlying skew because the estimators have a specific consistency correction for contaminated normal data (Lark, 2000a). We therefore had to transform data of events 1 and 5 before further processing.

For each event we constructed a Gaussian random field by unconditional simulation. The simulated values of fields 1 and 5 were back transformed to ensure that the fields reflected the structure of the original data. In a final step we contaminated the fields with outlying values of the respective event (Fig. 1). For an in-depth description of the construction of the fields we refer to Zimmermann et al. (2010) and Zimmermann and Zimmermann (2014). The fields have an edge length of 100 m, a grid unit of 0.1 m and hence consist of 10^6 data points.

A closer look at the simulated fields (Fig. 1, Table 1) reveals that they comprise a large span of spatial structures which is reflected in the variation of the nugget-to-sill ratio and the effective range, respectively. Relatively strong spatial structures and short autocorrelation distances characterize fields 3, 4 and 5. Accordingly, these fields have nugget-to-sill ratios <25% and effective ranges of around 3 m. In contrast, fields 1 and 2 feature somewhat weaker spatial structures but a comparatively long autocorrelation distance. Finally, field 6 features a pure nugget structure and hence displays no spatial correlation.

2.2. Sampling methods

For our study we tested the influence of the extent, sampling design, sample size, and methodology on the estimation of the variogram. In this section, we describe the selection of the extent, sampling design and sample size. In Sections 2.3 and 2.4 we deal with the variogram estimation methods.

To assess the influence of the extent, we employed three plot sizes with edge lengths of 25 m, 50 m, and 100 m, respectively. The largest plots comprised the entire simulation fields. For the smaller plots we arbitrarily chose the lower left corner of the simulated fields.

For the analysis of the sampling design we tested random sampling (R) in addition to three regular designs: a regular grid (G), a