Authors comments to Reviewer # 2:

Commments to the alledged „limitation of using only 1-D variograms“. The following has been added to the *revised* MS:

We here introduce the variographic approach mainly for the cases of 1-D and 2-D as a means of characterising the heterogeneity in the X-Y plane. Compared to the typical major variability in the Z-direction of soil depth profiles (soil horizons, layers, geological formations), the linear (1-D) or 2-D heterogeneity *within* soil horizons is significantly smaller, although this is exactly the kind of heterogeneity the present study aims at controlling. Contrary to depth profile zonation a.o. the within-horizon 1-D and 2-D heterogeneity complies with the requirements of both TOS and geostatistics, i.e. spatial heterogeneity can be modelled variographically w.r.t. a physically meaningful *average level* (the inherent *stationarity assumption* in geostatistics), e.g. Fig.s 2-5. It is not meaningful to apply variographic characterisation on measurement series which contain discontinuous shifts, upsets or other disrupt, level changes, as is the prime characteristicon of soil depth zonations. The geostatistical tradition of modelling 2-D patterns based on projection onto a 1-D transect can also be debated. In the present context all *isotropic* 2-D heterogeneity patterns can be characterised comprehensively by a *randomly selected* 1-D direction (transect). In all sampling operations there should preferentially always be some sort of random selection involved, unless compelling geo-science reasons exists for choosing a direction related to the genesis of the specific heterogeneity met with, e.g. choosing a 1-D transect either along a dominant plow direction.

Rational for a central Roman Square etc. The follwing has been added:

The experimental design allows comparison of the small-scale and large-scale variability. All profiles can for example be directly compared with the level and variation at the small-scale experiment, by the pertinent mean ±2 SD. This is just for visual orientation however and not to be confused with the nugget effect, a much more general characterisation of the small(est) scale variability pertaining to below lag = 1, summing up and averaging this information for all the sample pairs in the transect.

Comments re. PCA yielding an *average* range vs. individual, and specifically, the *minimum* range. The following has been added:

The results from the present study show that for well-mixed sandy soil it is recommended to sample locations with less than 2.5 meters inter-distance in between, preferentially smaller. It is necessary to conduct a similar variographic pilot experiment in order to outline the relevant scale-heterogeneity characteristics for other soil types, which unavoidably will tend to show more irregular spatial heterogeneity patterns – each principal soil type will in principle be characterised by a specific range, but there is a further caveat. Each analyte may in fact display its own, more or less specific range, as witnessed above, as well as by a plethora of studies in the literature. When controlling the spatial heterogeneity is of the essence, the logical solution is to design the sampling according the the analyte with the *smallest* range, i.e. the most heterogeneously distributed analyte – this will by necessity also take care of all other analytes with higher ranges. If emphasis is on sampling costs (a not totally unlikely alternative scenario that may, or may not clash with other requirements of which only one really matters though: representativity) it is a comforting thought that all analytes are measured on the same final aliquot. By carefully optimising the primary field sampling according to the principles presented here, all analytes will be measured with the same, optimal relevance, indeed w.r.t. the same representativity. If sampling is done right from the start, there are no extra costs – while the opposite is a very different case, as should be abundant clear.

Comments re. small vs. large scale dependency a.o.

The results from the present study show that for well-mixed sandy soil it is recommended to sample locations with less than 2.5 meters inter-distance in between, preferentially smaller. It is necessary to conduct a similar variographic pilot experiment in order to outline the relevant scale-heterogeneity characteristics for other soil types, which unavoidably will tend to show more irregular spatial heterogeneity patterns – each principal soil type will in principle be characterised by a specific range, but there is a further caveat. Each analyte may in fact display its own, more or less specific range, as witnessed above, as well as by a plethora of studies in the literature. When controlling the spatial heterogeneity is of the essence, the logical solution is to design the sampling according the the analyte with the *smallest* range, i.e. the most heterogeneously distributed analyte – this will by necessity also take care of all other analytes with higher ranges. If emphasis is on sampling costs (a not totally unlikely alternative scenario that may, or may not clash with other requirements of which only one really matters though: representativity) it is a comforting thought that all analytes are measured on the same final aliquot.

Specific soil types and/or other analytes will in principle display different ranges and nugget effects, and hence our call for systematic deployment of the *variographic pilot experiment*, from which can be derived all necessary information for designing an optimal sampling plan e.g. identifying the analyte with the smallest range (for significantly correlated analytes). For the case of well-mixed soil components, a general PCA-approach for modelling a whole set of variograms may be useful in addition to individual analyte consideration.

Without this type of information, experimental fate study work is essentially devoid a valid basis as regards interpretation, scale-up and scientific generalisation of the experimental results back to the field scale.

Primary - and sub-sampling relationships to TOS. The following has been added:

The primary sampling in this study (200-300g) was specifically intended to correspond to current sampling traditions in the soil and microbiology communities, so as to be as relevant as possible. We wanted to show the inherent deficiency in using a standardised sample size. In other studies efforts have been made to TOS-optimize each individual field sample, for example with respect to the famous “Gy’s formula”, from which control over the so-called Fundamental Sampling Error is often sought, while also controlling GSE and the incorrect (bias-generating) sampling errors. However, in the present study it is a major point to outline how the variographic approach a.o. lead to a procedure with which to characterize the magnitude of the total sampling-plus-analytical error and thus to be warned of the need to control (better) <u>all</u> the inherent sampling errors which are affected by both theDistributional Heterogeneity as well as the Constitutional Heterogeneity. The paper refer to the international standard DS 3077 (2013) for a comprehensive introduction to all aspects of sampling error reduction vs. lot/material heterogeneity in general and to Kardanpour et al. (2105b) in particular regarding these present variographic studies.

All sub-sampling steps were carried out with stringent attention to all TOS‘ principles for representative mass reduction. This has been described in minute detail in:

Kardanpour, Z, Jakobsen, O.S. & Esbensen, K.H. (2015b) Counteracting soil heterogeneity sampling for environmental studies (pesticide residues, contaminants transformation) - TOS is critical. Proceedings 7.th World Conference on Sampling and Blending (WCSB7), p.205-209.

“Why did subsampling not involve some sort of comminution”

In this study the sandy soil samples did not need comminution (because of the well-sorted, small grain size).