Focus maps: a means of prioritizing data collection for efficient geo-risk assessment

Massimiliano Pittore

Helmholtz Center Potsdam - German Research Center for Geosciences, Potsdam, Germany

Article history
Received October 29, 2014; accepted February 8, 2015.
Subject classification:
Seismic risk, Computational geophysics, Geo-risk assessment, Vulnerability, Data collection, Sampling.

ABSTRACT

Efficient Disaster Risk Reduction (DRR) management constantly calls upon high-quality information to be collected or updated for vulnerability monitoring and risk assessment. This process is often resource- and time-intensive, which many economically developing states (including most Central Asian countries) can seldom afford. In this paper, we introduce the concept of focus maps *as a useful tool to quantify the spatial probability of sampling. Focus maps allow for a data collection prioritization scheme to be put in place, enabling the realization of optimized spatial sampling which assigns a higher priority to locations where the need for high-quality information is greater. In practice, smaller samples can be drawn with the same (or better) resulting accuracy of the estimates, resulting in a more efficient use of time and resources. The factors that affect such a spatial sampling scheme include the usual components of risk assessment (hazard, exposure, vulnerability) where available, as well as other potentially critical factors, such as the extent and quality of previously collected data. The practical application of the proposed approach to the case of Central Asia will be exemplified and discussed.*

1. Introduction

Many countries are highly vulnerable to geohazards such as earthquakes, landslides and floods. In order to assess and quantify the risk arising from these natural threats, relevant reliable and up-to-date information on the exposed assets and the vulnerability of the involved communities is paramount for efficient Disaster Risk Reduction (DRR). Unfortunately, when little information is available - as is the case in many Central Asian countries - it has to be collected and integrated, which often proves to be a very time- and resource-intensive task. The collection of data *in-situ*, for instance, usually involves multiple survey teams, each composed by one or more persons, traveling in often difficult conditions to remote places.

It is thus important to devise prioritization strategies which would allow the practitioners to decide where to focus the most efforts and resources, in order to achieve the optimal trade-off between the requirements of DRR, the available resources and the constraints of the specific applications.

In this paper we propose a methodological scheme to integrate different layers of vulnerability-related information into a single map which we refer to as a *focus map*. Focus maps serve a two-fold purpose:

1) to provide a visual, intuitive representation of the hot-spots of an area with respect to end-users' interests or concerns, meaning the potential risk/loss arising from one or more natural hazards,

2) to allow for a consistent estimation of the desired density of information collection and sampling, conditional on the available data (the indicators), in order to precisely and efficiently drive the information updating process.

The application of focus maps therefore allows the implementation of an iterative risk assessment by alternating data collection and integration in a more efficient way.

Focus maps extend the concept of natural disaster hot-spots, already proposed in the literature (e.g., landslides, Nadim et al. [2006]; multiple hazards and risks, Dilley [2005]), in order to explicitly address the collection and integration of risk-related information into a continuous process where spatio-temporal indicators evolve over time. This involves not only following the natural variation of exposure, vulnerability and hazards, but also considering the evolution of the spatial extent and quality of the knowledge used to assess risk.

Our basic assumptions, throughout this paper, are as follows:relevant, reliable data are to be collected in the field with the ultimate purpose of assessing risk, generating models and/or validating previous inferences;

- we acknowledge the existence of several indicators, which alone or jointly will affect the data collection activities;

- these indicators may functionally affect the risk (e.g., level of hazard, or human exposure), or directly impact upon the data collection itself (physical accessibility, cost, specific dangers);

- the specific (analytic) functional relationship between these indicators and the data collection activities is not known in advance (usually because the dependency model is complex/non-linear/unknown and not enough information is available to properly constrain it);

- end-users have a basic understanding of the relative weight of the available indicators (that is, in what proportion they affect the risk or the data collection itself).

Let us remark that with the term *end-users*, we refer to a group of individuals (or institutions) which need to draw conclusions or plan further assessment activities in order to understand the possible impact of one or more natural hazards on the exposed communities. This group might include decision- as well as policy-makers, and, for instance, civil protection authorities.

From the formal point of view, the discussion surrounding these issues could be framed within the broader perspective of building composite indicators, which has been a theme in the literature (see, e.g., Nardo et al. [2005] for a comprehensive review). Composite indicators are often used to generate proxies for risk estimation at global and regional scales (see for instance Dilley [2005], Dao and Peduzzi [2004]).

Composite indicators can be useful for policy analysis, and have proven useful in benchmarking country performance. Moreover they seem easier to interpret than finding a common trend in many separate indicators. On the other hand, composite indicators can send misleading policy messages if they are poorly constructed or based on uncertain data, and could lead end-users (especially policy makers) to draw simplistic conclusions [Saisana and Tarantola 2002, Nardo et al. 2005].

Indicator	**Unit of measure**	**Type**
population density per unit area	counts	real ≥ 0
average GDP per unit area	currency	real ≥ 0
seismic hazard as PGA (Peak Ground Acceleration) with exceeding probability 10% on 50 years	cm/s^2	real ≥ 0
flood inundation scenario with a recurrence interval of 100 years	meters	real ≥ 0
susceptibility with respect to landslides	susceptibility index, dimensionless	real $\geq 0, \leq 1$

Table 1. Examples of vulnerability and risk-related indicators.

We acknowledge the fact that often the available data is not sufficient to generate reliable models by statistical inference, but end-users often have empirical evidence of the indicators and their relative importance. Moreover, there could be factors which are directly connected with the efficiency of the data collection activities and have no direct influence on the assessment of risk, but still have to be considered in a pragmatic approach to end users' needs (for instance, the availability and quality of legacy data or logistic constraints).

We propose a shift in this paradigm, by focusing our attention on the factors that evidently affect the collection of high-quality data, and on the optimization of the data collection itself, which would bring the information needed to better constrain a realistic model, rather than generating imperfect inferences of risk that rely on multiple proxies.

Nonetheless, the fundamental building blocks or processing stages of composite indicators must be considered, namely:

- normalization,
- aggregation,
- weighting.

The inclusion of these stages is a natural necessity of the new paradigm we are introducing, where the final goal is not the assessment of risk, but the implementation of a joint density of probability of sampling (for information collection) which is conditionally dependent upon the indicators themselves. Since such a joint conditional probability is difficult to achieve in practice, when few data is available, we simplify the problem by considering instead an aggregated probability based on the combination of several components. Each selected indicator is therefore substituted by a spatial probability of sampling given the indicator itself.Within this framework, the *normalization stage* is interpreted as a functional mapping which defines the probability of a certain location to be sampled given the value of the indicator itself in that specific location. The *aggregation* in our case is the process of combining several probabilities following different schemes. In this paradigm, we can rely for instance on the existing literature about probability aggregation (see for instance Clemen and Winkler [1999], Ranjan [2009], Allard et al. [2012]). The necessity of a *weighting* naturally arises from the selected aggregation approach.

In the following section, the interpretation of focus maps within a statistical sampling perspective will be further detailed, with several examples provided that focus on Central Asia.

2. Statistical interpretation of focus maps

In this section an introduction to the theoretical background to focus maps is provided, along with a description of how they are related to the sampling task. Let:

$$D_i = D_i(x,y) \in [0,\infty], (x,y) \in G \tag{1}$$

be a set of 2-dimensional datasets defined on a continuous or discrete geographic extent G. For the sake of simplicity, we suppose that the geographic extent is common to all the indicators (same extent, origin and projection).

Each dataset, which can be described by a 2-dimensional map, represents an indicator which D_i is relevant either to the estimation of risk or for the information collection itself. We also suppose that all indicators are real and not negative. A few examples of possible indicators are listed in Table 1.

2.1. Probability mapping (normalization)

Let us define a new map set as:

$$S(D_i) = P(S \mid D_i) \in [0,1] \tag{2}$$

representing the probability of each location to be selected for data collection, given the value of the particular indicator D_i (at the considered location). The value $S(D_i)$ is intuitively a quantification of the importance of the indicator for the estimation of risk, since we are interested in collecting data starting from locations whose contribute to the overall risk evaluation is higher. The conditional probability $P(S \mid D_i)$ can be defined by a simple mapping on [0,1], which is mathematically equivalent to a normalization.

Several approaches can be followed to realize such mapping. Typically this is realized by a combination of functional forms and truncation. A few examples are reported in the following and summarized in Table 2.

Mapping operator	Description	Formula
minmax	linear mapping	$\frac{D_i - \min_G(D_i)}{\max_G(D_i) - \min_G(D_i)}$
logarithmic	logarithmic mapping	$\beta_0 + \beta_1 \ln(D_i + \epsilon)$
quadratic	quadratic mapping	$\beta_0 + \beta_1 D_i^2$
inv-logit	inverse logistic regression mapping	$\frac{e^{\beta_0+\beta_1 D_i}}{1+e^{\beta_0+\beta_1 D_i}}$
log-square	squared logarithmic mapping	$\beta_0 + \beta_1 \ln(D_i + \epsilon)^2$

Table 2. Examples of possible mapping operators.

The mapping defined by Equation (3) (*minmax*), for instance, implements a simple stretching of the input indicator's values. This is simple and can be suitable in many cases, but may be affected by the presence of outliers.

$$S(D_i) = P(S \mid D_i) = \\ = \frac{D_i - min_G(D_i)}{max_G(D_i) - min_G(D_i)} \qquad D_i \in (-\infty, \infty) \tag{3}$$

In this case, a truncation based on rejection bounds can be used to improve the robustness of the mapping. Rejection bounds are therefore defined based on a quantile interval (e.g., from 5% to 95%) and all indicator values outside the considered rejection bounds are trimmed to the boundary values, therefore excluding the tails of the distribution which would have dominated the resulting probability $P(S \mid D_i)$.

The mapping in Equation (3) can also be biased by distributions with a significant dynamic range. In this case, a logarithmic mapping can be used, such as, for instance:

$$S(D_i) = P(S \mid D_i) = \\ = \beta_0 + \beta_1 \ln(D_i + \epsilon) \qquad D_i \in (0, \infty) \tag{4}$$

or a squared logarithmic (*log-square*):

$$S(D_i) = P(S \mid D_i) = \beta_0 + \beta_1 \ln(D_i + \epsilon)^2 \tag{5}$$

where ϵ represents a small coefficient to account for null values, and β_0 and β_1 are generic parameters. The normalization in general can be described by a more complex mapping function, for instance to better account for the signal's dynamics over the value ranges of interest:

$$S(D_i) = P(S \mid D_i) = \\ = \frac{e^{\beta_0+\beta_1 D_i}}{1+e^{\beta_0+\beta_1 D_i}} \qquad D_i \in (-\infty, \infty), \beta_i \in \Re \tag{6}$$

Equation 6, for instance, refers to a mapping based on an inverse *logit* functional, with two degrees of freedom. This is equivalent to defining a univariate logistic regression described by the coefficients β_i. In this case, the interpretation of the coefficient can be intuitively explained. β_0 defines the "baseline" probability, when the value of the indicator is equal to zero. The coefficient β_1 defines the sign of the conditional dependence (if positive, there will be a positive correlation between the indicator's values and the resulting probability, and conversely, if it is negative). The absolute value of β_1

indicates how much the final probability changes for a given increment of the indicator's value.

The values of the parameters in Equations (4), (5) and (6) can be either specified by the user, or computed from the data by linear regression, specifying a small set of sampling probabilities. The choice of the most appropriate mapping function is important and should be carefully evaluated. For example, a simple descriptive statistic of the indicator (e.g. quantiles analysis) may often already provide useful insights into the particular mapping to be realized.

2.2. Probability pooling

Ideally, we would estimate the joint probability:

$$SJ(x,y) = P(S \mid D_1, D_2, ..., D_n) \tag{7}$$

representing the posterior probability for each spatial location (x,y) to be selected for data collection, given the value of all the considered indicators D_i (at the considered location) as covariates. Unfortunately, this is often not possible, or is exceedingly complicated. In order to define the joint probability defined by Equation (7), the full interdependencies among the covariates and the posterior probabilities should be known, which is very rarely the case.

We can therefore approximate the conditional probability with a suitable pooling operator *PG* defined as:

$$\begin{aligned} &SJ(x,y) \approx PG(P(S \mid D_1), ..., P(S \mid D_n)) = \\ &= FM(D_1, ..., D_n) \end{aligned} \tag{8}$$

We hereby define a focus map as this approximation of the joint probability of sampling. To create a focus map, we must therefore first specify a *mapping* for each indicator, as explained in the preceding section, then indicate the most appropriate *pooling* operator. Several pooling operators have been proposed in the literature for probability aggregation. The most common operators can be grouped into two families:

- additive (and additive transformed) methods;
- multiplicative methods.

2.3. Additive pooling operators

Additive methods refer to linear mixture models, and are related to the union of events (logical OR). The single conditional sampling probabilities are simply weighted and summed. Therefore, if the weights are positive, the contribution of the different layers is cumulated. A significant contribution by one single layer will hence tend to drive the final result. The use of additive (linear) pooling is suggested when the relative importance of the considered layers needs to be considered. Linear pooling is defined as:

$$\begin{aligned} &PG(P(S \mid D_0), P(S \mid D_1), ... \; P(S \mid D_n)) = \\ &= \sum_{i=0}^{n} w_i P(S \mid D_i) \end{aligned} \tag{9}$$

where the constraint:

$$\sum w_i = 1, w_i > 0 \forall i$$

applies.

A different approach to linear pooling is represented by the *beta-transformed linear pooling* [Ranjan and Gneiting 2010, Allard et al. 2012]:

$$PG = H_{\alpha,\beta}\left(\sum_{i=0}^{n} w_i P(S \mid D_i) \right) \tag{10}$$

This method has been proposed in literature [Ranjan and Gneiting 2010] to overcome some the limitations of the standard additive approaches, and will not be discussed in the present work.

2.4. Multiplicative pooling operators

Multiplicative methods intuitively relate to the intersection of events (logical AND). A multiplicative approach tends to emphasize the spatial locations where all the involved indicators forecast a higher probability of sampling, and penalizes the spatial locations where at least one of the involved indicators has a low sampling probability. The impact of the use of such pooling on the resulting focus map is often significant, and underpins a specific usage pattern. Multiplicative approaches are in fact particularly useful for prioritizing the data collection, by giving immediate relevance only to those locations which collect the greater "consensus" among the considered indicators.

The most simple and useful method for multiplicative aggregation is represented by *log-linear pooling*:

$$\begin{aligned} &\ln PG(P(S \mid D_0), P(S \mid D_1), ... \; P(S \mid D_n)) = \\ &= \ln Z + \sum_{i=0}^{n} w_i \ln P(S \mid D_i) \end{aligned} \tag{11}$$

with the following constraint:

$$\sum w_i = 1, w_i > 0 \forall i$$

on the weights, or alternatively:

$$PG \propto P_0^{1 - \sum_{i=1}^{n} w_i} \prod_{i=1}^{n} P_i^{w_i} \tag{12}$$

In this case, we can drop the restrictions on the positivity and bounds of the weights since the result of the pooling will be always bounded in the [0,1] interval. Furthermore, we also have to note that the use

of weights greater than 1 further enhances the selective effect of the pooling.

3. Benefits of focus maps in developing sampling strategies

Within the framework of data collection and survey design, a focus map can be interpreted as a representation of the spatially-varying inclusion probability, which is the probability for each point of the area frame to be selected and included in a sample. Focus maps are thus a means to exploit auxiliary information to achieve higher statistical efficiency on the one hand, and a reduction of the total cost of the survey on the other.

In this section, the application of focus maps to probability sampling will be explored and discussed, in comparison with a standard sampling design based on Simple Random Sampling (SRS) (see Cochran [1977], Wang et al. [2012], Stevens and Olsen [2004] for a comprehensive overview of spatial sampling methods). In order to investigate the effectiveness of the proposed approach, a simplified test case is considered, and a stochastically generated spatial distribution set is used as a possible realization of realistic target distributions.

Let us suppose that in a hypothetical area-frame, representing the geographic boundary of the area of interest, the total risk arising from a certain natural phenomenon (such as an earthquake) is to be estimated. For the sake of simplicity, we define the *simplified risk* (R) as

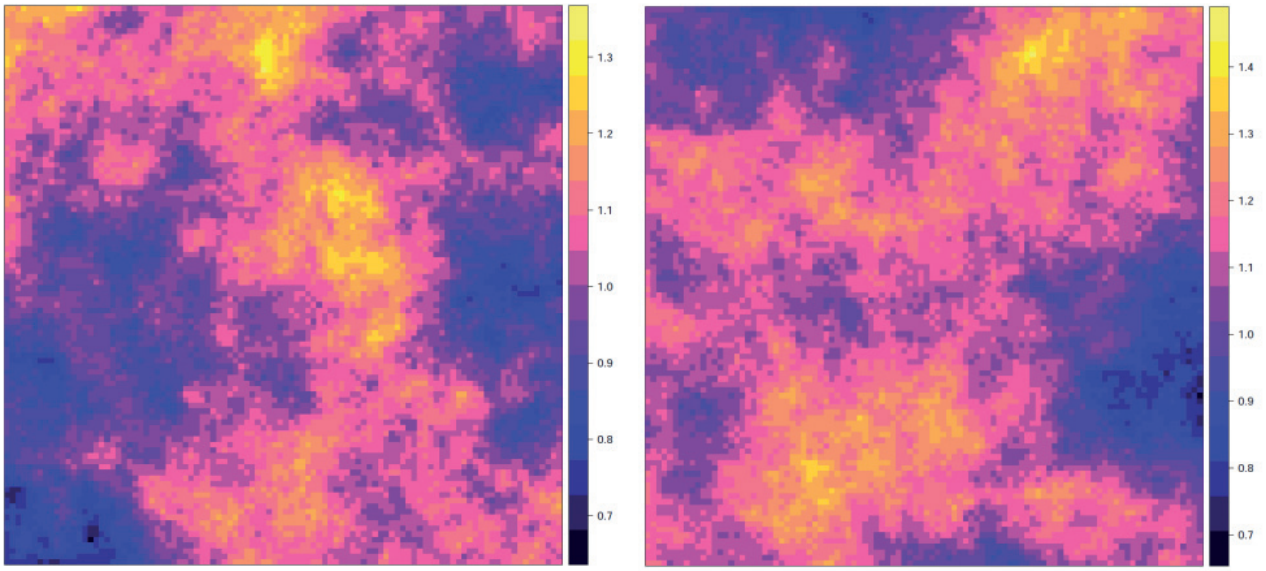

Figure 1. Simulated spatial distributions of hazard (left) and exposure (right) over a common area frame, gemerated by stochastic realization of Gaussian Random Fields.

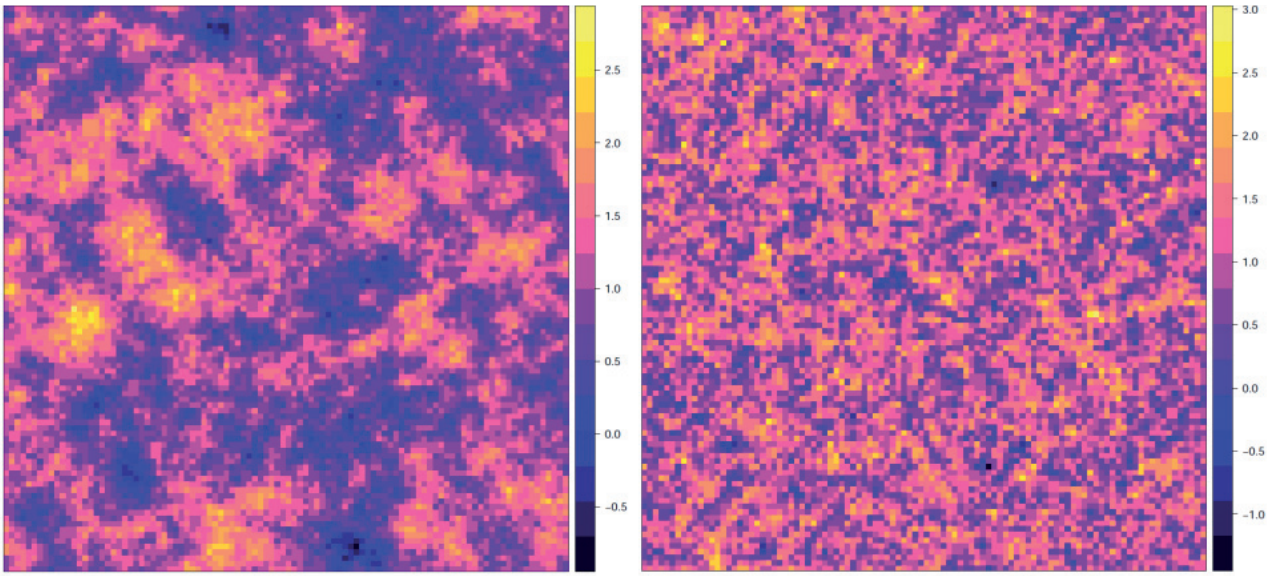

Figure 2. Simulated spatial distribution of (physical) vulnerability with two different levels of spatial autocorrelation, gemerated by stochastic realization of Gaussian Random Fields. The distribution on the left side has a higher coefficient of spatial correlation with respect to the distribution on the right side.

the multiplicative combination of three components: namely hazard (H), exposure (E) and vulnerability (V):

$$R(x) = H(x) \cdot E(x) \cdot V(x) \tag{13}$$

where x represents a specific location within the considered area.

We will define the three components in Equation (13) as scalar spatial distributions modelled by Gaussian Random Fields, that is, probabilistic distributions with a non-zero degree of spatial autocorrelation [Christakos 1992]. This assumption is based on the observation that often spatial distributions representing physical quantities show a certain amount of spatial autocorrelation, that is, locations closer in space will likely exhibit similar characteristics. This phenomenon is often referred to as the Tobler law [Tobler 1970, Miller 2004].

Figures 1 and 2 depict the simulated distributions, respectively, of hazard H and exposure E (the latter representing, for instance, the distribution of population over the considered area), and vulnerability V (meaning, for instance, the physical vulnerability of the residential buildings). The parameters of the semivariogram describing the spatial auto-correlation of the stochastically generated Random Gaussian Fields are reported in Table 3. All distributions are defined over an area frame defined by 10,000 locations (indexed by a 100×100 grid). The particular choice of the semivariogram parameters has been made in order to simulate a realistic case. The spatial scale of the area frame is such that the distributions of seismic hazard and population exhibit a certain degree of autocorrelation. Also, the physical vulnerability of the buildings inhabited by the people is supposed to be partly autocorrelated, but a smaller spatial range can be expected, since the factors contributing to the physical vulnerability are subject to higher degrees of spatial variability.

Figure 2 shows two different realizations of a vulnerability distribution with respectively higher and lower spatial autocorrelation.

We suppose that the estimate of *total simplified risk*

Parameter	Sill	Range	Type
hazard	0.05	100	exponential
exposure	0.025	50	exponential
vulnerability	0.25	5	exponential
vulnerability (alternative)	0.25	1	exponential

Table 3. Autocorrelation parameters for the stochastic generation of the test distributions.

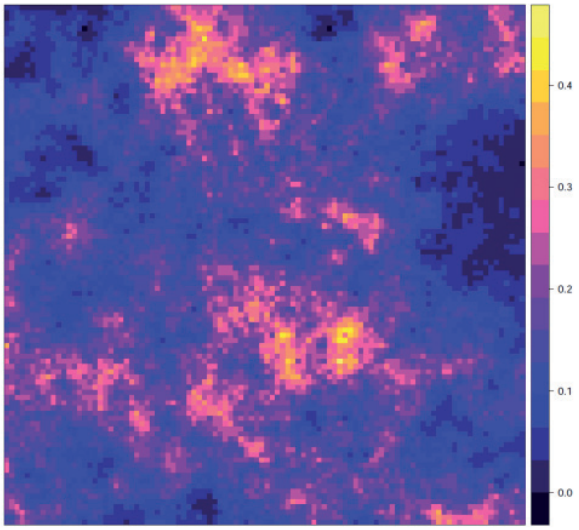

Figure 3. Distribution of simplified risk obtained as product of the simplified distributions of hazard, exposure and vulnerability as described in Equation (13).

R_T is of interest to us. The total simplified risk is thus defined as:

$$R_T = \sum_A R(x) \tag{14}$$

where A represents the considered area frame.

The spatial distribution of simplified risk, as defined by Equation (14), is shown in Figure 3. The vulnerability distribution depicted in Figure 2 (left) has been used.

The vulnerability distribution is supposedly not known in advance, therefore it has to be sampled in order to correctly estimate the risk. Since the actual estimation of seismic vulnerability can be a time-expensive and burdensome process, such a survey has therefore to be carefully planned.

The simplest solution would be to use a sampling design based on a one-stage survey with Simple Random Sampling (SRS). To simulate several realizations of such a type of survey, we extract several samples of the same size from the area frame A and for each sample estimate the total risk R_T.

A summary of the statistics for the SRS estimation is provided in Table 4. The estimator's mean refers to the mean of the values provided by the estimator for

Sampling design	True value	Estimator mean	Estimator coeff. of variation	Estimator bias (%)
SRS (ss=100)	1464.26	1466.12	3.57	0.13
STR (ss=100)	1464.26	1464.05	0.013	0.015
PPS (ss=100)	1464.26	1464.06	1.71	−0.014

Table 4. Comparative results of estimations based on different sampling designs. SRS=Simple Random Sampling, PPS=Probability Proportional to Size, SS=Sample Size. 100 samples without replacement, each replicated 1000 times and averaged.

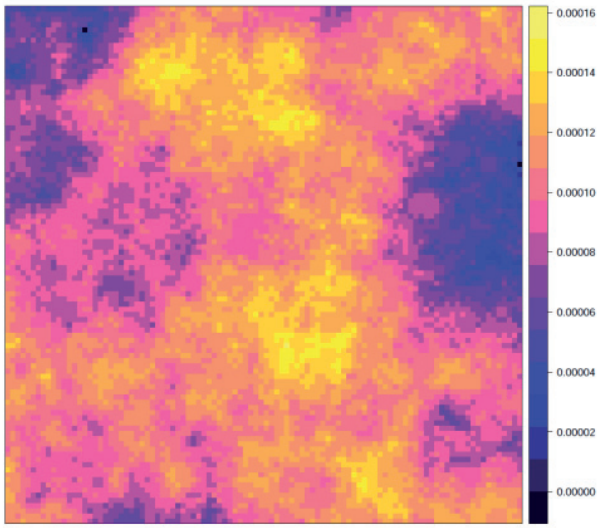

Figure 4. Resulting focus map obtained by the multiplicative pooling of the simulated spatial distributions of hazard and exposure depicted in Figures 1 and 2. A simple min-max mapping, and an equal weighting scheme have been used.

1000 replicas (independent draws with replacement) of SS = 100 samples each. Every sample refers to a possibly different spatial location within the considered area, and the inclusion probability is constant at every location (and equal to the inverse of the number of locations). The estimator coefficient of variation (CV) is defined as the variance of the estimator divided by its mean. The estimator bias is defined as the average difference between the estimated value and the true value. The CV provides information on the precision of the estimator, while the bias defines its accuracy. A good estimator should therefore be both accurate and precise.

The disadvantage of SRS is that it does not consider any auxiliary information to improve the efficiency of the survey. Often, information which can be directly or indirectly related to the features of interest (in this case the total simplified risk) can be often collected. In this case, for instance, we can suppose that both hazard and exposure are known (or can be estimated) and can be exploited to improve the efficiency of the survey.

We can therefore compute a focus map *fm* by pooling together the two distributions, using a log-linear approach:

$$fm = \frac{e^{w_1 \cdot \log H + w_2 \cdot \log E}}{\sum_A e^{w_1 \cdot \log H + w_2 \cdot \log E}} \tag{15}$$

where the weights are balanced ($w_1 = w_2 = 0.5$) and the two distributions have been normalized using a *min-max* linear mapping with no rejection bounds. The resulting distribution, shown in Figure 4, is also normalized in order to correctly represent an inclusion probability.

The computed focus map not only can be considered a proxy of simplified risk, but it can also be used to drive a more efficient sampling strategy. As an example, a Probability-Proportional-to-Size (PPS) sampling design can be implemented using the focus map as the unequal probability of inclusion. In Figure 5 two samples drawn according to the two different methodologies are shown. It is interesting to note that, even though the samples appear quite similar, the two sampling designs are profoundly different. Furthermore, the results in Table 4 show that the PPS sampling based on the computed focus map achieves a higher precision then SRS.

A further test is presented which compares the two sampling designs with increasing sample size. The resulting standard deviation of the *total risk* estimator (fil-

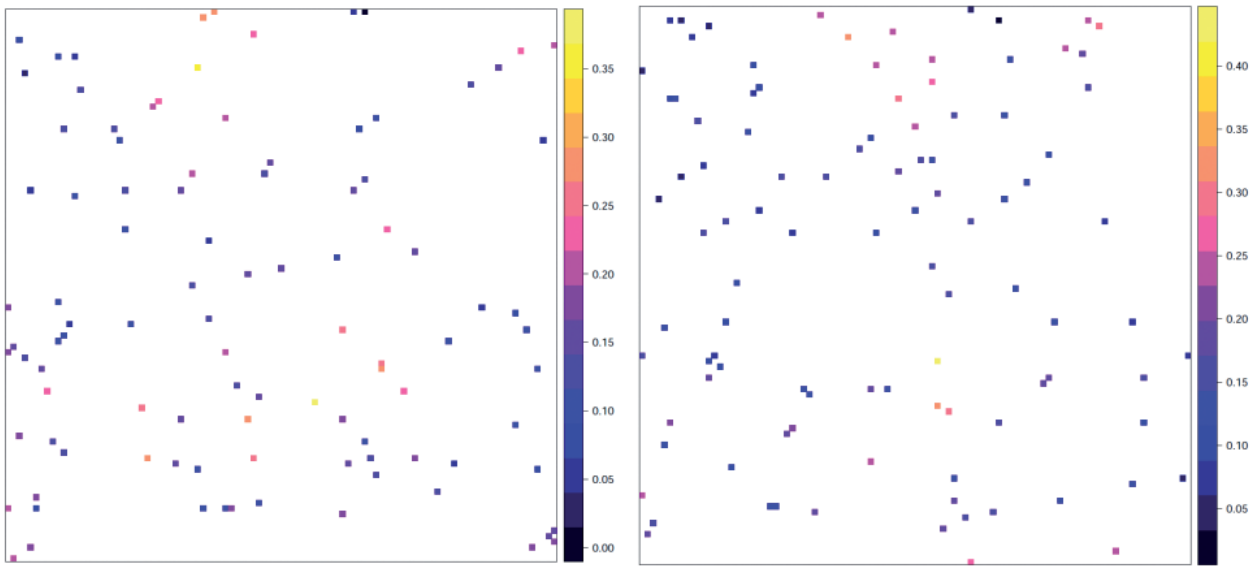

Figure 5. Two spatial samples drawn according to the two different methodologies: PPS (left) based on focus maps, and SRS (right). The location of points selected in the samples are colored according to the value assumed by the underlying target distribution (simplified risk, assumed as unknown). The two samples appear similar, but have been generated by two very different approaches.

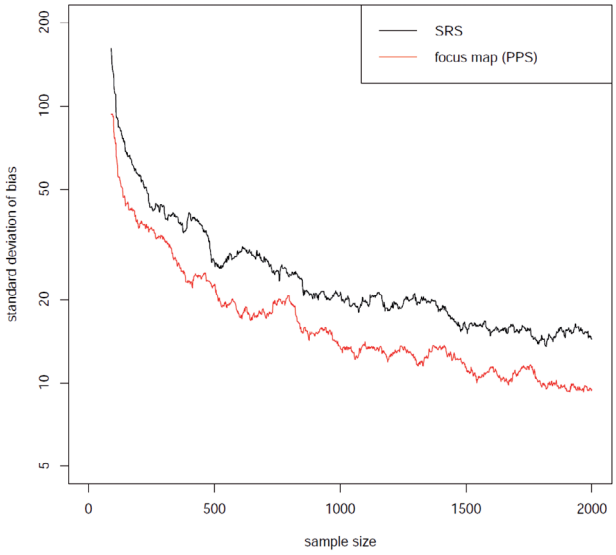

Figure 6. Standard deviation of the bias computed for the two different sampling approaches: PPS based on focus maps and SRS. The standard deviation is computed over a running buffer with 90 point, and refers to samples of increasing size. The PPS approach has a lower bias with respect to the purely random one.

tered through a 90-points mobile average) is shown in Figure 6.

The estimator variance is systematically lower for the sampling design based on the focus map. The PPS sample design is particularly useful when total estimates are desired, since they are more sensitive to higher values. In other cases, for instance when average estimates are needed, information has to be collected so as to cover a broader range of values assumed by the target variable. In those cases, other sample designs could be better suited. In the following, a *stratified sample* approach (STR) exploiting the focus map will be exemplified. Following this approach, the area frame is subdivided into smaller areas (strata) where the target variable is deemed to have an homogeneous value, and samples are independently drawn from each stratum [Cochran 1977].

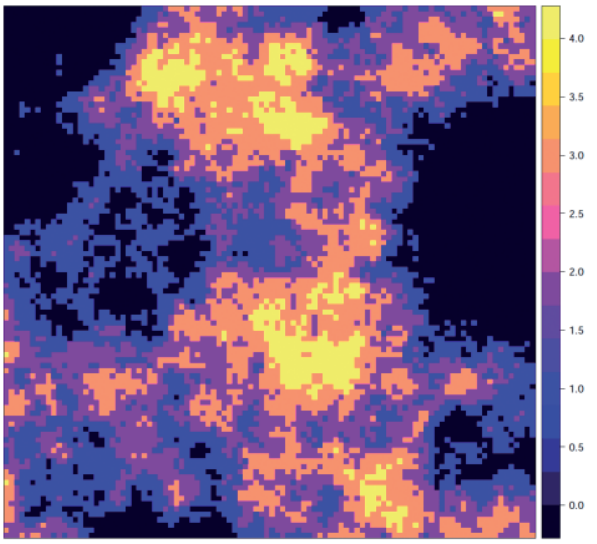

Figure 7. A spatial stratification of the area frame is shown. The stratification is obtained by selecting areas corresponding to different quantiles of the focus map as original distribution. In particular, the quantiles (0.25, 0.5, 0.75, 0.95) are used to generate 5 different strata. The first three strata have 2500 units each, the last two strata have respectively 2000 and 500 units.

A spatial stratification is obtained by quantization [Gersho 1977] of the computed focus map over the area frame. This is based on the assumption that the variable of interest, in this case the simplified risk, is relatively homogeneous in the computed strata, where a SRS sampling design can then be applied. This method has the advantage, with respect to the PPS sampling design, that areas with both high and low risk will be sampled, therefore yielding a more accurate estimate of properties such as average and proportion. On the other hand, such a sampling design is characterized by a greater sparseness of sample points, thus resulting in less efficient survey implementation on the field.

In Figure 7 a stratification of the area frame is shown. The stratification is obtained by selecting areas corresponding to different quantiles of the computed focus map. In particular, the quantiles (0.25, 0.5, 0.75, and 0.95) are used to generate 5 different strata. The first three strata have 2500 units each, the last two strata have respectively 2000 and 500 units.

For each stratum, a SRS with proportional allocation is selected (with 1% sample size). The results, averaged on 1000 replicas, are also listed in Table 6. We note that the STR estimator based on the focus map exhibits much better performances than the SRS sample design, and comparably also better precision with respect to the PPS based estimator.

4. Example: ranking and selecting populated places for risk assessment

In order to exemplify the application of focus maps, we consider the task of selecting the locations to be prioritized on a regional scale for further risk assessment. The geographical scope covers the countries of Uzbekistan, Tajikistan and Kyrgyzstan in Central Asia. The settlements layer we intend to sample is composed of 10,522 locations, representing most of the populated places in the selected region (see www.geonames.org). We consider the following input layers:

- spatial distribution of population (Landscan™ 2012);

- spatial distribution of seismic hazard, in terms of MSK-64 [Grünthal 1998] macroseismic intensity with a exceedance probability of 10% in 50 years [Ullah et al. 2014];

- by combining the places with a higher density of population with the locations exhibiting higher levels

50%	75%	95%	99%
0	3	63	954

Table 5. Percentiles of the Landscan distribution of population for all of Kyrgyzstan. The distribution has an extended dynamic range, exceeding 42dB, with 95% of the grid population values below 63, and 99% of the grid cells with less than 1000 inhabitants.

Sampling probability $P(S \mid D)$	Indicator value (population count in each grid-cell)				
	1	100	500	1000	10,000
	0.001	0.1	0.3	0.5	0.9

Table 6. Percentiles of the Landscan distribution of population for all of Kyrgyzstan. The distribution has an extended dynamic range, exceeding 42dB, with 95% of the grid population values below 63, and 99% of the grid cells with less than 1000 inhabitants.

of hazards, it is possible to obtain a representation of the risk hot-spots in the region. A focus map can therefore drive this process in a straightforward way, without conflicting with the need of a more comprehensive assessment of risk.

The two input layers are first normalized using two different mappings.

We consider as input the Landscan™ 2012 distribution of population over the area of interest. The distribution has an extended dynamic range, exceeding 42dB, with 95% of the grid population values less than 63, and 99% of the grid cells with fewer than 1000 inhabitants (see Table 5), while the maximum value of the population in a grid cell of the considered area is 18,965. The distribution is therefore very skewed.

In a first stage, the input layers are subject to suitable mappings, in order to obtain individual sampling probabilities. The parameters of the mapping operators (except *minmax*) can be easily determined by linear regression using the probability sampling model specified in Table 6. In Figure 8, the conditional probability distributions based on four different normalizations are shown on a smaller scale. The first (upper left corner) is a *minmax* mapping with (5%-95%) rejection

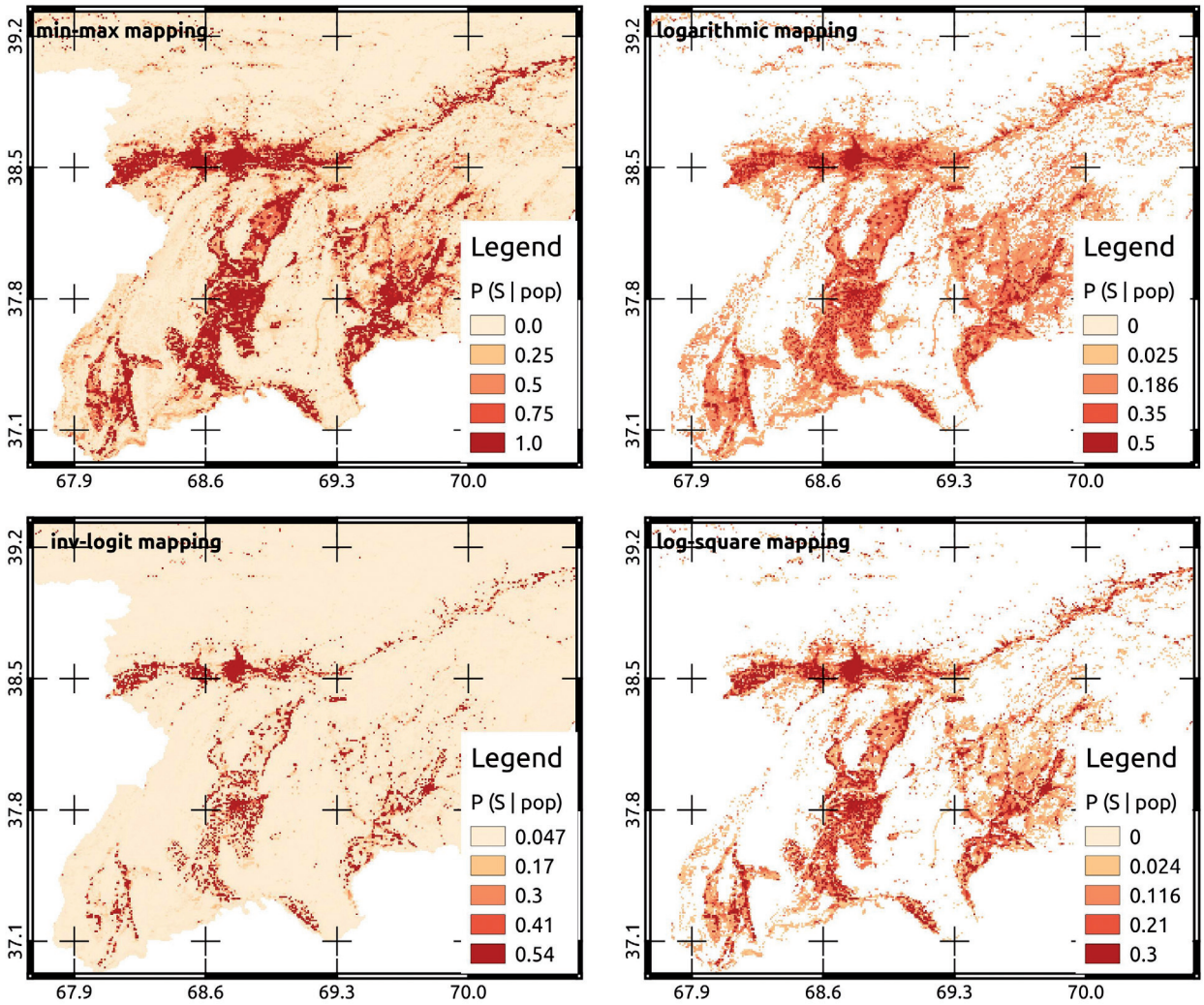

Figure 8. Comparison among different types of mapping for an indicator describing the distribution of population for a portion of the region of interest. Upper left: minmax, upper right: logarithmic, lower left: inv-logit, lower right: log-square. The mapping parameters have been determined by linear regression using the sampling probability associations specified in Table 6.

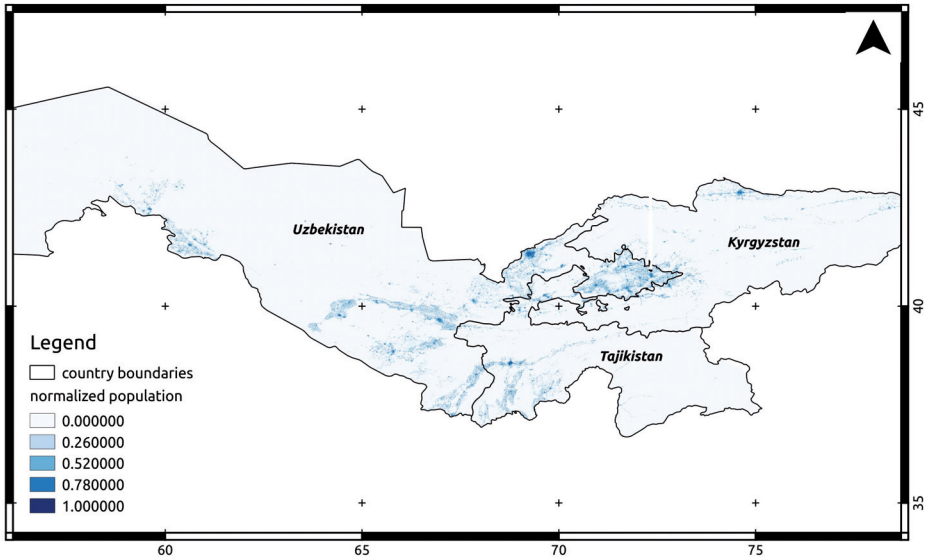

Figure 9. Normalized distribution of inhabitants in the selected region, including the Central Asian countries Kyrgyzstan, Tajikistan and Uzbekistan.

bounds, the second (upper right corner) is a pure logarithmic mapping, the third (lower right) is a *log-square* mapping, while the fourth (lower left corner) is defined by a *inv-logit* transformation. The figure shows the impact of the different normalizations on the conditional probability of sampling. We note how both the basic min-max normalization and the inverse-logit do not change radically the probability distribution. As a matter of fact, the inverse-logit slightly decreases the sampling probability for grid cells with low population values, and saturates for higher values. On the other hand, the *logarithmic* and *log-square* normalizations have an equalizing effect on the histogram of the sampling probability. Grid cells with small population values are assigned a much higher sampling probability with respect to the linear and inv-logit case.

The population layer is therefore mapped according to the log-square operator defined in Table 2 and calibrated with the sampling model listed in Table 6 (the normalized layer is shown in Figure 9). This mapping equalizes the dynamic range of the input layer, which in the region tends to be dominated by a few large settlements.

The seismic hazard, expressed as macroseismic intensity EMS-98 [Grünthal 1998] with exceedance probability of 10% in 50 years, is subject to a quadratic mapping

Sampling probability $P(S\|D)$	**Indicator value (seismic hazard, macroseismic intensity EMS-98)**				
	I	III	V	VIII	X
	0.0001	0.01	0.3	0.7	0.99

Table 7. Associations between hazard values and desired sampling probability values for the determination of the mapping parameters. Hazard is expressed in terms of EMS-98.

(see Table 2) calibrated with the sampling model listed in Table 7. This mapping, shown in Figure 10, accounts for the non-linearity of the effects related to changes in the macroseismic intensity. Moreover, it also allows us to substitute an ordinal variable (the macroseismic intensity), for which no metric is defined, with a numerical one, representing a conditional sampling probability.

A log-linear operator is chosen to pool the two sampling probability functions. In Figure 11A, the result of the multiplicative pooling, with equal weighting, is displayed. In Figure 11B, an alternative focus map is obtained by choosing an equal weighting scheme with both weights equal to one. The result is a strongly selective focus map.

By using an unequal weighting scheme, and assigning a weight of 0.9 to the hazard layer, and 0.1 to the population layer, the population layer acts as a mask to filter out regions not inhabited and with lower levels of seismic hazard (see Figure 11C).

The focus map obtained with equal weighting (see Figure 11A) is used to rank the 10,000+ populated places considered in the region. For each of the three countries considered, the places with focus map values greater than 0.9 indicating the highest sampling probability according to the considered layers and the respective weighting, are selected for further investigation.

The selected locations are displayed in Figure 12. According to the estimated ranking priority, more in-depth risk assessment activities should be carried out in the Ferghana Valley, as suggested by the selected locations in the Andijon and Namangan provinces, Uzbekistan, and in the southern part of the Kathlon province in Tajikistan. Together, these three provinces account for around 7 million inhabitants (according to the 2012 census), more than 10% of the overall population of

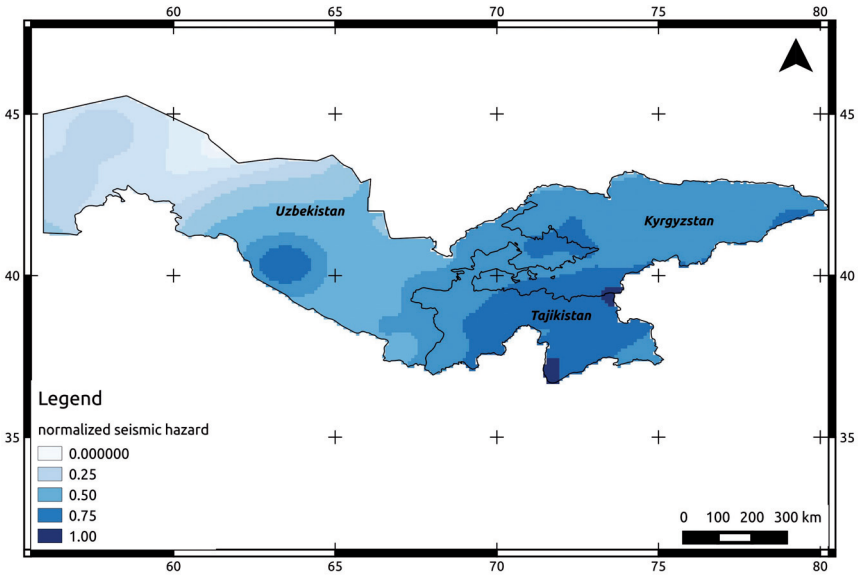

Figure 10. Normalized distribution of seismic hazard in the considered region. The input layer is defined as of macroseismic intensity (MSK-64) with exceedance probability of 10% in 50 years.

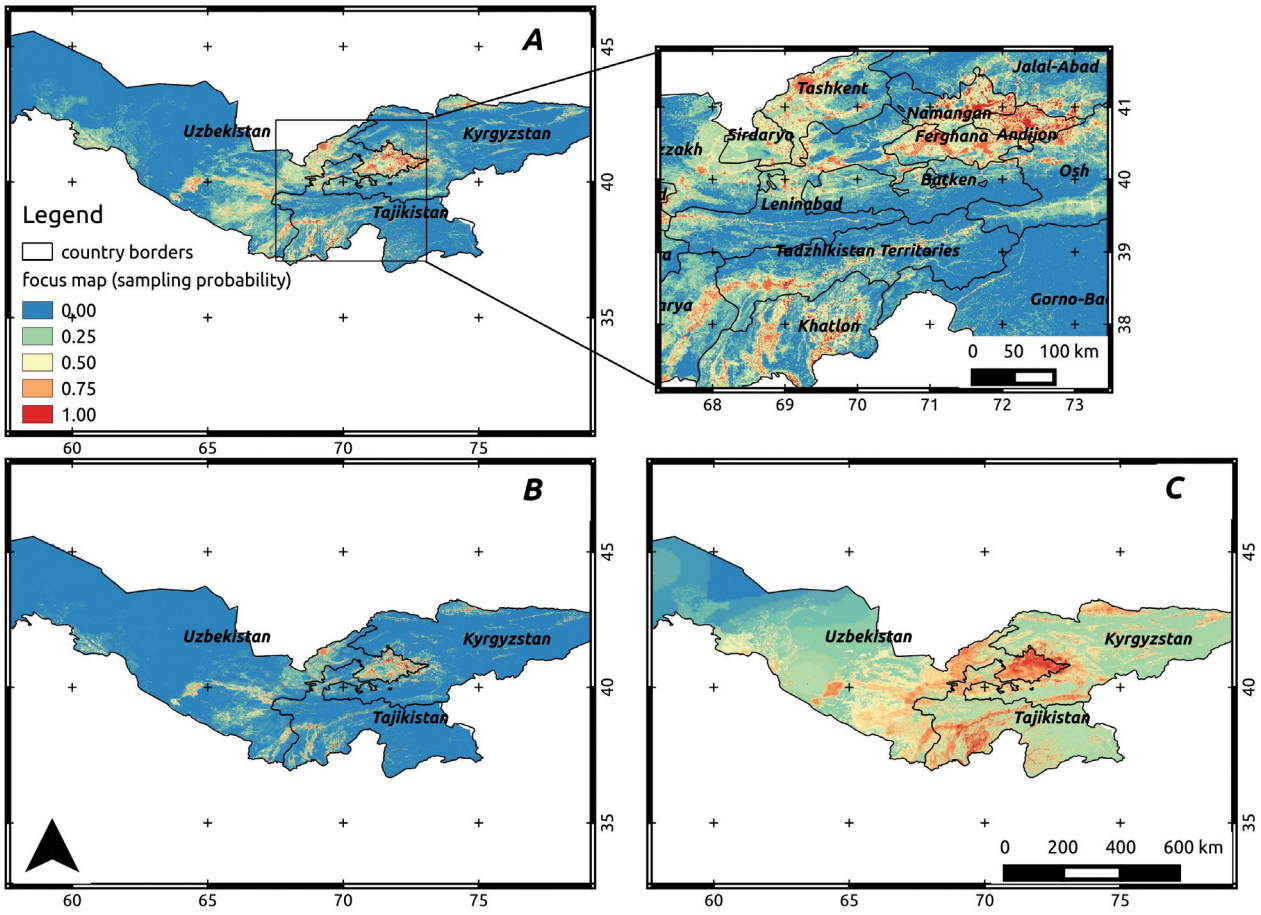

Figure 11. Comparison between focus maps obtained by loglinear pooling with different weighting schemes. A: equal weighting (w=0.5) B: equal weighting (w=1) C: unequal weighting (w=0.9 for seismic hazard layer. The inset shows a close-up of the area with a higher sampling probability, focusing on the Ferghana Valley.

Central Asia, and mostly in rural environments. Moreover, the town of Andijon, in the Homonym province, was severely damaged in 1902 by an earthquake that resulted in approximately 4500 casualties [Kondorskaya et al. 1982]. Attention should be also paid in Kyrgyzstan to the towns of Jalal-Abad and Kyzyl-Kiya, respectively in the Jalal-Abad and Batken provinces, as well as in several sparse locations in the Gorno-Badakshan region in Tajikistan. Let us remark, however, that such a ranking analysis does not take into account the expected seis-

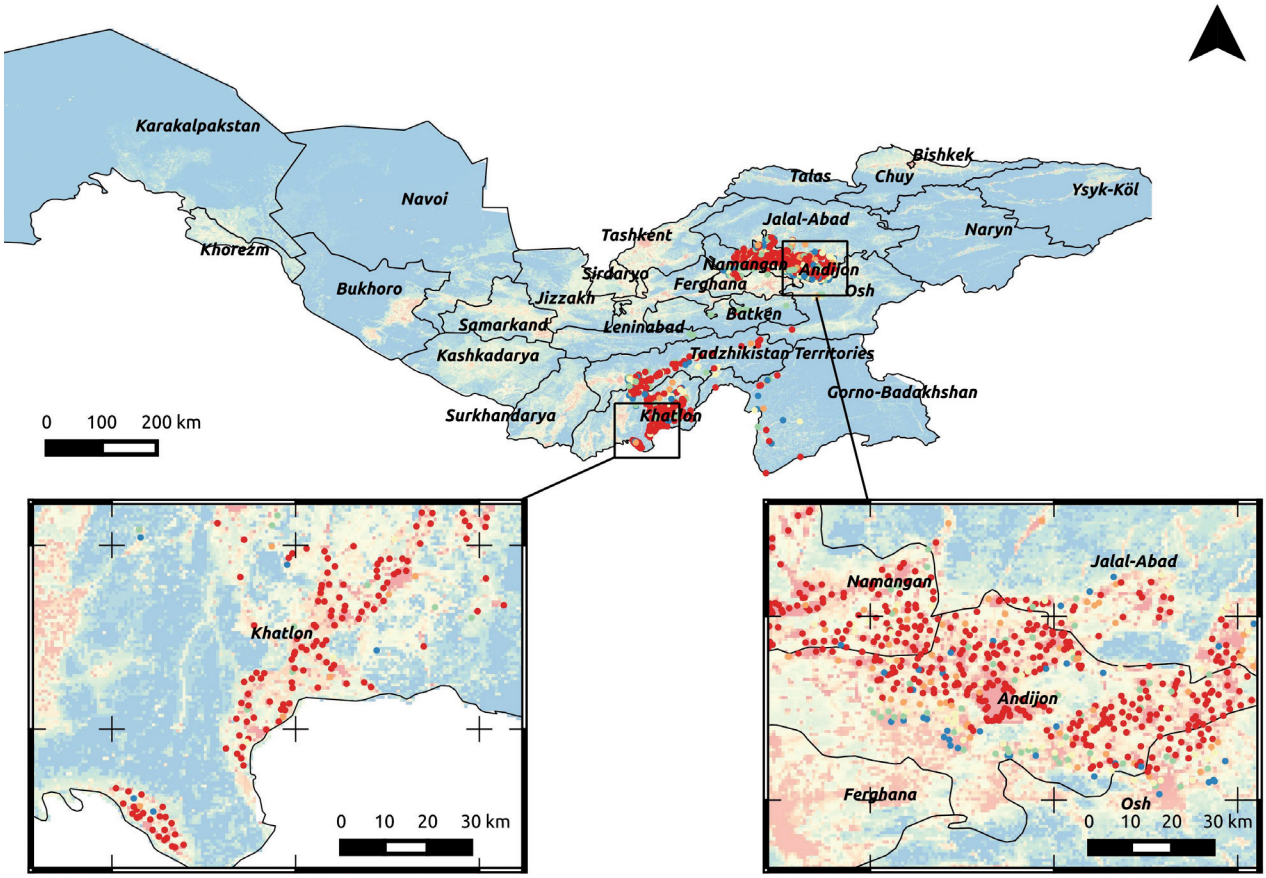

Figure 12. Sampling of settlements in the region of interest for further risk assessment. The focus map is used to rank all populated places in the region. Only the settlements with sampling probability greater than 0.9 (for each of the three considered countries) are selected.

mic vulnerability in the target area, which should be indeed the focus of *in-situ* investigation suggested by this preliminary evaluation.

5. Conclusions

In the present paper we have introduced a geo-statistical methodology which allows for the implementation of optimized spatial surveys aiming at risk characterization. We refer to this scheme as a focus map, and it provides a two-fold advantage:

1) an intuitive representation of the spatial hotspots of a set of composite indicators, representing the end-users' interests or concerns. In this framework, focus maps can be correlated with the risk arising from different natural hazards, with less chance of them being misinterpreted as actual risk distributions;

2) focus maps are described by spatial distributions which can be employed as inclusion probabilities in the realization of efficient spatial sampling approaches.

The actual implementation of focus maps is based on two basic steps, namely *mapping* and *pooling*. *Mapping* normalizes the input indicator's values into a probability range [0,1], according to a user defined scheme. Several different mapping approaches are proposed to either smooth the effect of outliers in the input distribution, or to account for very skewed distributions that exhibit a high dynamic range. *Pooling* refers to the combination of probability distributions (a result of a previous mapping phase) into a final probability distribution. Two main approaches for combining probabilities have been proposed, that is the implementation of either additive (linear) or multiplicative (log-linear) pooling schemes. Additive approaches provide a basic combination of input layers, and can be useful when independent layers have to be considered to drive field-based data collection. For instance, probabilistic modeling of different hazards (e.g., earthquakes and floods) can be linearly combined to optimize the collection of common vulnerability indicators. On the other hand, multiplicative approaches acknowledge the mutual functional interdependence among the input layers and are more suitable for prioritization schemes. For instance, by log-linearly pooling hazard and exposure -related probability layers, a higher sampling (inclusion) probability will be given to spatial locations where both indicators are relevant, while strongly penalizing those locations where at least one of the input indicators is negligible.

The advantage of sampling designs based on focus maps has been investigated in a simplified framework. In particular, probability-proportional-to-size (PPS) and stratified (STR) samplings based on focus maps have been compared with simple-random-sampling (SRS)

approaches. As the results suggest, the use of focus maps based on auxiliary information related to the feature of interest allows for a more efficient sampling to be implemented. In practice, smaller samples can be drawn with the same (or better) resulting accuracy of the estimates, resulting in a more efficient use of time and resources.

While the pooling of different indicators has been already presented in the literature within other contexts, the novelty of the proposed approach lies in the definition of a consistent paradigm for geo-information collection and integration, rather than on the implementation of generic risk-proxies. Furthermore, the statistical interpretation of mapping and pooling in terms of spatial sampling probability suggests further research lines towards the realization of iterative sampling and integration schemes, as well as the realization of more sophisticated algorithms for data aggregation and analysis.

Acknowledgements. This work was supported by the SENSUM (framework to integrate Space-based and in-situ sENSing for dynamic vUlnerability and recovery Monitoring, Grant agreement no. 312972), PROGRESS (Georisiken im Globalen Wandel) and EMCA (Earthquake Model Central Asia) projects. The Author is grateful for useful advice and suggestions from Dr. Daniele Ehrlich, and would like to thank Dr. Kevin Fleming for his kind revision of the English. Landscan™ 2012 High Resolution global Population Data Set copyrighted by UT-Battelle, LLC, operator of Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the United States Department of Energy. The United States Government has certain rights in this Data Set. Neither UT-Battelle, LLC nor the United States Department of Energy, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of the data set.

References

Allard, D., A. Comunian and P. Renard (2012). Probability aggregation methods in geoscience, Mathematical Geosciences, 44, 545-581.

Christakos, G. (1992). Random field models in Earth sciences, San Diego, Academic Press, 474 pp.

Clemen, R., and R. Winkler (1999). Combining probability distributions from experts in risk analysis, Risk Analysis, 19, 187-203.

Cochran, W.G. (1977). Sampling Techniques, 3rd edition, Wiley.

Dao, H., and P. Peduzzi (2004). Global evaluation of human risk and vulnerability to natural hazards, In: Proc. EnviroInfo 2004 conference, 435-446.

Dilley, M. (2005). Natural disaster hotspots a global risk analysis, Washington, D.C., World Bank.

Gersho, A. (1977). Quantization, IEEE Communications Society Magazine, 15, 16; doi:10.1109/MCOM.1977.1089500.

Grünthal, G., ed. (1998). European macroseismic scale 1998 (EMS-98), European Seismological Commission, Subcommission on Engineering Seismology, Working Group "Macroseismic Scales", Cahiers du Centre Européen de Géodynamique et de Séismologie, 15, Luxembourg, 99 pp.

Kondorskaya, N.V., N.V. Shebalin, Y.A. Khrometskaya and A.D. Gvishiani (1982). New catalog of strong earthquakes in the USSR from ancient times through 1977, World Data Center A for Solid Earth Geophysics, Report SE-31, Boulder, Col., USA, 608 pp.

Miller, H.J. (2004). Tobler's first law and spatial analysis, Annals of the Association of American Geographers, 94, 284-289.

Nadim, F., O. Kjekstad, P. Peduzzi, C. Herold and C. Jaedicke (2006). Global landslide and avalance hotspots, Landslides, 3, 159-173.

Nardo, M., M. Saisana, A. Saltelli and S. Tarantola (2005). Tools for composite indicators building, European Commission, Ispra.

Ranjan, R. (2009). Combining and Evaluating Probabilistic Forecasts, University of Washington.

Ranjan, R., and T. Gneiting (2010). Combining probability forecasts, Journal of the Royal Statistical Society Ser. B, 72, 71-91.

Saisana, M., and S. Tarantola (2002). State-of-the-art report on current methodologies and practices for composite indicator development, Report 20408, European Commission-JRC, Italy.

Stevens, D.L., and A.R. Olsen (2004). Spatially Balanced Sampling of Natural Resources, Journal of the American Statistical Association, 99, 262-278; doi:10.1198/016214504000000250.

Tobler, W.R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region, Economic Geography, 46, 234-240.

Ullah, S., D. Bindi, M. Pilz, L. Danciu, G. Weatherhill, E. Zuccolo, A. Ischuk, N.N. Mikhailova, K. Abdrakhmatov and S. Parolai (2014). Probabilistic seismic hazard assessment for Central Asia, Annals of Geophysics, 58 (1), S0103; doi:10.4401/ag-6687.

Wang, J.-F., A. Stein, B.-B. Gao and Y. Ge (2012). A review of spatial sampling, Spatial Statistics, 2, 1-14; doi:10.1016/j.spasta.2012.08.001.

Corresponding author: Massimiliano Pittore,
Helmholtz Center Potsdam - German Research Center for Geosciences, Potsdam, Germany; email: pittore@gfz-potsdam.de.