Journal of Natural Disaster Science, Volume 40, Number 2, 2020, pp14-25

Visualizing Data Saturation Process in Mapping Site Amplification of Earthquake Ground Motions

Anirban CHAKRABORTY* and Hiroyuki GOTO**

*Department of Urban Management, Graduate School of Engineering, Kyoto University
**Disaster Prevention Research Institute, Kyoto University

(Received: Sep. 10, 2019 Accepted: Jan. 21, 2020)

Abstract

Seismic hazard maps play an important role in earthquake disaster risk reduction. The availability of spatial data is crucial to generate these maps that plot the spatial distribution of hazard potentials to emphasize spatial differences. The past few decades have seen an exponential increase in the availability of geospatial data. However, we cannot ascertain whether the amount of available data is sufficient, and we have no guidelines to draw the maps based on the available data consistent with the data accumulation. In this study, we address these issues in terms of data visualization techniques. Using information theory, we propose a parameter that measures the incremental information gain as maps are updated with new data over time. Data saturation occurs as the proposed parameter approaches zero. The concept is applied to a case study area in the Furukawa district of Japan where earthquake data has been collected over 7 years from 31 seismometers in a dense seismic array. Convergence in site amplification maps generated over different observation periods conclude that the mapping in Furukawa district is approaching data saturation and from the viewpoint of information theory, the current operation may be terminated.

Keywords: data visualization; disaster risk reduction; seismic hazard map; site amplification; uncertainty

1. INTRODUCTION

Earthquake disasters are significant events causing large-scale destruction to human society. In order to reduce the extent of earthquake damage, possible scenarios should be assessed taking into account the spatial variation of the seismic risk. Earthquake ground motion amplifications are the main cause of these spatial differences (e.g., Kawase, 1996) and have recently been summarized in seismic hazard maps.

In order to generate seismic hazard maps that plot the spatial distribution of earthquake ground motion amplifications or other hazard potentials, the availability of spatial data is important. In the past few decades, advancement in data collection, e.g., high-resolution remote sensing, monitoring sensor networks, etc., has increased the availability of spatial data considerably (Lee and Kang, 2015).

However, it is usually not clear if the amount of available data is sufficient to extract the desired information for the physical process. More data collection has been attempted if the amount of data is determined to be insufficient. This concept is based on the idea that the goal of the hazard maps is to plot accurate information using sufficient data, which can contribute to disaster reduction.

On the other hand, in practice, we need to plot the maps based on only the available data. In such a situation, there are two essential questions: (1) how do we determine whether the amount of available data is sufficient, and (2) how do we draw the maps consistent with the data accumulation?

In this study, we address these issues in terms of data visualization techniques. Visualization is an important aspect to help the user directly understand the data saturation. The objective of this study is to visualize data saturation in mapping. We introduce KL divergence increments, based on information theory (Kullback and Leibler, 1951), that measures the incremental information gain as we update a map. Data saturation or sufficiency is reached when no more incremental information gain is observed even after adding new data to a map. In the literature, papers addressing the issue of data saturation in spatial mapping are rare. Some papers have addressed the issue of optimal sampling, however, the focus is on assessing the grid layout rather than the optimal amount of data (Hughes and Lettenmaier, 1981; Wang et al., 2012; James and Gorelick, 1994). Past research in other fields has seen some papers addressing the issue of data saturation (Chaudhuri et al., 1998; Guest et al., 2006; Fusch and Ness, 2015). However, many of them are qualitative in nature and none of them considers data uncertainty in their formulations and hence the question on reliability remains unanswered.

In this study, we consider Uncertainty Projected Mapping (UPM) (Chakraborty and Goto, 2018) as the mapping tool. Unlike conventional mapping (e.g., ordinary kriging), which does not consider the record to record variability and uses only a single averaged value for each site, UPM projects the data uncertainty on the map resolutions and adds statistical significance to the estimated values at the sites. Also, as we shall see in this paper, UPM has an interesting characteristic of being dependent on the number of observations. This characteristic is crucial to enable use of our proposed parameter in measuring the information gain as more and more data is added to a map. The novelty of the idea lies in involving data saturation with the spatial resolution.

In this study, we examine a methodology to visualize and quantify the excess or deficiency of data in mapping earthquake ground motion amplifications. In the next section, we discuss the UPM model and introduce the incremental KL divergence parameter to quantify data saturation. In Section 3, we introduce a numerical experiment to discuss how the parameter can help in quantifying and visualizing data saturation in spatial maps. In Section 4, we apply our methodology to a real earthquake site amplification dataset from a case study area in Japan and discuss how the results can help us decide when to stop collecting more data.

2. METHODOLOGY

2.1 Uncertainty Projected Mapping (UPM)

Mapping is a popular visualization tool applied to understand a spatial process. Kriging (Matheron, 1963) is a very useful tool to spatially interpolate data based on spatial variations. However, most of the conventional visualization techniques assume that the data is free of uncertainty and uses only the mean (μ_j) at site j (Brodlie et al., 2012). Many researchers believe that displaying uncertainty on maps could lead to better decisions (Harrower, 2003). Our motivation is to put more information (uncertainty) into the map resolutions and hence, in this study, we use UPM as the mapping tool.

UPM considers two uncertainties in a spatial process: record to record variability at a site j (σ_j) and site to site variability in the neighborhood of j (s_j). The record to record variability is the same as the standard deviation at site j. There are two goals of UPM: (1) to project the information of σ_j into posterior estimates of μ_j at site j, and (2) to decrease the map resolution in zones of high σ_j. Using (1), it adds statistical significance to the estimated mean (μ_j) and hence, it can statistically explain the difference in values at the two sites. To achieve (2), a constraint relation is introduced such that (Chakraborty and Goto, 2018)

$$c = s_j \sigma_j \tag{1}$$

where c is a constant. Equation (1) relates the smoothness or roughness of map resolutions to σ_j , i.e., the record to record variability at site j. Generally, s_j , i.e., the site to site variability in the neighborhood of j, determines the form of the map resolutions. A low s_j value means low variability around j and hence, a smooth resolution. However, a high s_j means a rough variability around j and thus, a rough resolution. In UPM, we want the map resolutions to be smooth in zones of high σ_j. Equation (1) achieves this by constraining an inverse relation between σ_j and s_j . In zones of high σ_j we impose a low s_j and make the resolutions smooth. In zones of low σ_j, we impose a high s_j and make the resolutions rough. The constant c is unique to the model setting and the optimum value of the constant c is based on model evaluation as discussed later.

Thus, in UPM, the mapping follows conventional mapping (e.g., ordinary kriging) in areas with low record to record variability. However, as the record to record variability increases, the mapping becomes smooth and similar to the neighboring sites. This is the uniqueness of UPM. It can project the information of the uncertainty in the map resolutions.

Neighborhood is an important component in modelling UPM. In many cases, the spatial sites may not be uniformly spaced or there may be some missing sites where the values need to be estimated. So, in general, we create uniformly distributed sites (by adding missing sites, if necessary) so that ideally every site has the same number of neighbors.

UPM is based on a Bayesian hierarchical model (Banerjee et al., 2014). The unknown parameters μ,σ and s are assigned a prior distribution and estimated based on a posterior probability distribution using MCMC (Gilks and Spiegelhalter, 1995) algorithms. The plot of the μ thus obtained is called the UPM of the spatial variable. However, based on different c values, many different UPMs can be generated. In Chakraborty and Goto (2018) model evaluation was conducted using a k-fold cross-validation (Stone, 1974). However, it is computationally very expensive as the model data needs to be split into many parts. To avoid that splitting of the dataset into subsequent parts, in this study, we replace cross-validation with computationally faster Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010; Gelman et al., 2014). The c model with the minimum WAIC is considered as the best model. In this paper, all the UPM results come from the optimal c model.

2.2 ΔD_{KL} : Proposed parameter to quantify data saturation

The estimated mean (μ_j) and estimated record to record variability (σ_j) improve as more and more data is added to the mapping. In other words, maps evolve with the addition of data in time. In this study, we visualize and quantify this property of maps and decide the point of data saturation, which means that a further increase in data adds no more information to the map.

As we will see in Sections 3 and 4, UPM has a property of converging with conventional mapping (ordinary Kriging) as the number of data increases. To quantify this convergence in UPM, we use a parameter based on Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). KL divergence measures the degree of difference between two probabilistic distributions. It is usually defined as

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{2}$$

where P and Q are continuous random variables and p and q are the associated probability densities.
In this study, we define a quantity called incremental KL divergence (ΔD_{KL}) given by

$$\Delta D_{KL_{[N+\Delta N]}} = \sum_j D_{KL}\left(P_{[N],j} || P_{[N+\Delta N],j}\right) \quad (3)$$

where $\Delta D_{KL_{[N+\Delta N]}}$ is the D_{KL} between the probability distribution $P_{[N]}$ at the N observation data case and the probability distribution $P_{[N+\Delta N]}$ at the $N+\Delta N$ observation data case summed over the j sites. In the literature, such convergence measure has been proposed to evaluate the performance of numerical analysis (Goto and Bielak, 2008).

The parameter ΔD_{KL} measures the information gain as the maps are updated with more and more data in time. Data saturation occurs when ΔD_{KL} approaches zero, which means that no more spatial information is added even upon adding more data to the map. The uniqueness of the parameter ΔD_{KL} is that unlike conventional measures of data saturation, it also considers the data uncertainty in its formulation and hence adds a sense of reliability to the measurement.

It is difficult to define a ΔD_{KL} like parameter to measure data saturation in conventional mapping (ordinary Kriging). The reason can be explained based on the differences between the mapping characteristics of UPM and ordinary Kriging, which can be listed as follows: (1) Unlike UPM, ordinary Kriging needs to estimate the distribution of both mean (μ_j) and record to record variability (σ_j), separately. (2) Unlike UPM, no statistical dependences between the mean and record to record variability are incorporated in ordinary Kriging. (3) Unlike ordinary Kriging, the UPM maps vary with the amount of observation data. When there is less observation data, UPM maps are smooth. The ordinary Kriging maps, on the other hand, are rough even when there is less observation data. However, as the observation data increase, the UPM maps change and approach the conventional mapping. This change in UPM map characteristics with the amount of observation data helps quantify the convergence process.

3. NUMERICAL EXPERIMENT

3.1. Data

In engineering seismology, it is well established that alluvial deposits can significantly affect the amplitudes of the incident seismic waves (Trifunac, 1971). Past studies indicate with high confidence (low uncertainty) that the waves become highly amplified at around the center of an alluvial valley. However, there is a low confidence (high uncertainty) concerning the amplification characteristics at the boundary between the rock site and alluvial valley. This is because the incident angle and frequency contents are well affected.

For the numerical experiment, we create a hypothetical model of the wave amplification in and around an alluvial valley as a spatial process (Fig. 1). Sixty-three sites are considered in one dimension. The sites are all equally spaced except at the boundary between the rock site and alluvial valley where the sites are more densely spaced. The reason behind this is to properly model the change in uncertainty as one moves away from high uncertainty at the boundary to the low uncertainty regions.

Random wave amplification samples (mean values shown as black dots in Fig. 1) are artificially generated as observations at all these sites. Each observation sample refers to the wave amplification observed for incident seismic waves during one earthquake event. The random samples at each site follow a lognormal distribution. The known arithmetic means (μ) increase as a parabolic curve from the rock site to the center of the valley representing high wave amplification. The record to record variability (σ) is treated as the uncertainty, which is high in the boundary zone of the rock site and the alluvial valley and low otherwise. The extreme amplitude changes at the basin boundary occur as the incident angle and frequency contents are well affected at the basin boundary (Trifunac, 1971).

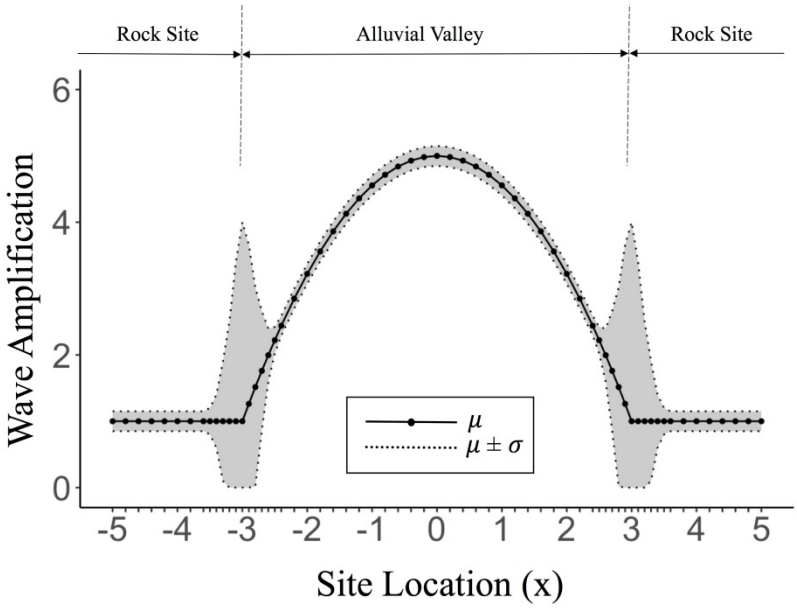

Fig. 1. Numerical experiment: A hypothetical model of wave amplification in and around an alluvial valley

3.2 Results

Figure 2a shows eight different cases with 8, 16, 32, 64, 128, 256, 512 and 1024 artificial random samples per site. Each succeeding dataset with higher earthquake events includes the preceding dataset with lower earthquake events. In all these cases, the gray circular dots show random samples, the blue line shows the Kriging mapping, and the red line shows the UPM.

For cases with a low number of observations (*N*), the UPM shows a smooth transition in the highly uncertain boundary zone between the rock site and alluvial valley, unlike the conventional mapping which is very rough and fluctuating. This smoothness is introduced by Equation (1). The boundary zone has a high σ_j. So, UPM makes the transition smooth by imposing a low s_j in the boundary zone. A low s_j means the μ_j values around site j do not vary much which makes the UPM smooth. However, in the low uncertainty areas including the center of the valley, UPM behaves like Kriging. UPM maintains the Kriging shape in areas of low σ_j by imposing a high s_j around j. A high s_j means the μ_j values around site j vary considerably and hence no smoothness is introduced in UPM in areas other than the boundary zone.

As the number of observations (*N*) increases, UPM starts to converge with Kriging. This change in the characteristics of UPM with the increase in the number of observations (*N*) is significant to gain an understanding of the population.

When the number of observations is low, there is less information for modelling and so the estimated model parameters are quite unstable. The conventional mapping for the low observation dataset when compared with the known mean values is erroneous in the high uncertainty zone. In such a situation, the smoothness introduced in the UPM in the highly uncertain boundary zone between the rock site and alluvial valley is a better representative of the physical process than the erroneous conventional mapping.

However, when the number of observations (*N*) is high, there is more information for modelling and so the estimated model parameters are stable. The conventional mapping for the 1024-observation dataset when

compared with the known mean values is almost the same. Due to increased data, error is also reduced in the high uncertainty region. It is very interesting to observe that the UPM now converges with the conventional mapping. This finding shows that UPM yields reliable results as compared to conventional mapping when less information is available and can be used to hint at data saturation as the number of observations increases.

Figure 2b shows the incremental KL divergence (ΔD_{KL}) with respect to the number of observations, calculated using Equation (3). Sites located at the edges are not included in the calculation of ΔD_{KL} because we want to discuss the results as an interpolation problem. At the edge, the values are estimated as an extrapolation problem. It is observed that ΔD_{KL} starts to converge as the number of observations increases. This indicates that the mapping on UPM reaches convergence and the data set is sufficient to extract the population statistics. Among them, we can set up the observation strategy to refer to the evolution of ΔD_{KL} through the UPM.

It is difficult to take a similar approach using conventional mapping (ordinary Kriging). In Fig. 2a, UPM gives a smoother transition under poor data information, which is very reasonable in the sense of appearance. However, the ordinary Kriging has an unreasonably rough transition under poor data information. When data information is richer, the UPM maps and the ordinary Kriging maps approach the hypothetical model of wave amplification in Fig. 1. This transition of UPM maps from reasonably smooth to rough mapping with the change in information quantity, allows the quantification of the convergence process.

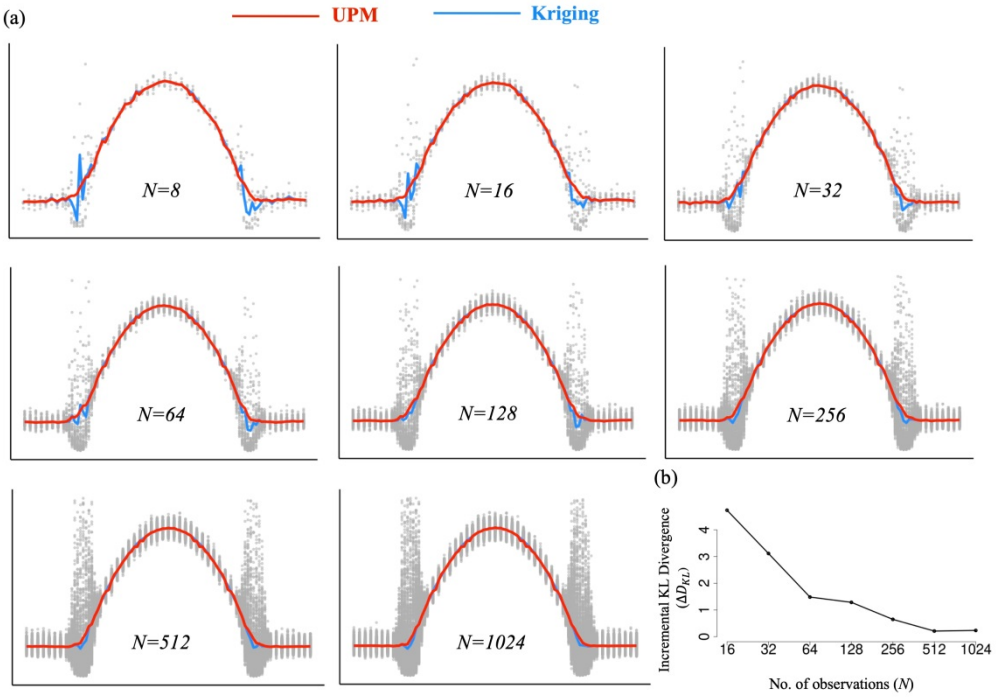

Fig. 2. (a) Evolution of UPM and Kriging maps for numerical experiment
(b) Plot of ΔD_{KL} vs N for the UPM maps

4. APPLICATION: REAL DATA

4.1 Data

The 2011 off the Pacific coast of Tohoku Earthquake caused heavy damage to life and property due to the tsunami and the strong ground motion. Severe damage occurred not only close to the shoreline, but also in areas further into the mainland. Furukawa district in Osaki City, Miyagi prefecture of Japan experienced severe damage in downtown residential areas (Goto and Morikawa, 2012). Significant spatial differences caused mainly due to ground motion amplification (site amplification) were observed even on the sub-kilometer scale. In the aftermath of the earthquake, a very dense seismic network for strong ground motions has been operated in Osaki city (Goto et al., 2012). Figure 3 shows the layout of the seismic array consisting of 31 seismometers in the significantly damaged area in Osaki city. The seismic array observation was jointly organized by Kyoto University, Tokyo Institute of Technology, Osaki city office and aLab Co., Ltd.

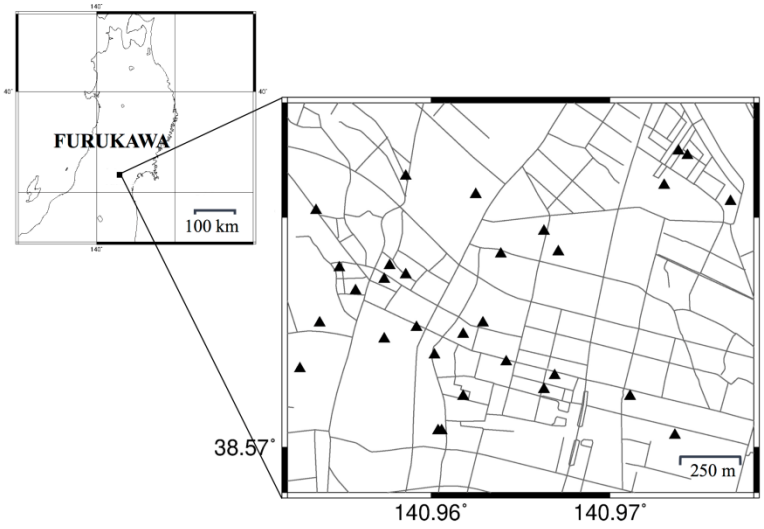

Fig. 3. Spatial distribution of seismometers (▲) in Furukawa district, Japan

In this case study, earthquake data collected over 7 years from 31 sites in the seismic array is used to generate a site amplification map of the area. A total of 176 earthquake events recorded between October 29, 2011 and September 19, 2018, were used for the analysis. These earthquake events are mostly aftershocks from the 2011 off the Pacific coast of Tohoku Earthquake and include all recorded events in the above-mentioned period without any restriction on the amplitude threshold or source location condition. The average peak ground acceleration (PGA) of the recorded events ranges from 6 gal to 119 gal. The availability of observation data varies with the sites. To study the convergence process, 6 datasets were created using groups of 8, 16, 32, 64, 128 and 176 earthquake events. Each succeeding dataset with higher earthquake events includes the preceding dataset with lower earthquake events.

The mapping parameter in this case study is a factor of site amplification observed at site *j* during an earthquake event. It is defined as the logarithmic ratio of observed peak ground acceleration (PGA) or peak ground velocity (PGV) at site *j* to the spatial average calculated over all the available sites during one earthquake event. The PGA and PGV are calculated from the vector sum of the EW component and NS component of the earthquake record. To generate a UPM map of the site amplification, the dataset was comprised of 431 sites with 31 measurement sites from the seismic network and 400 missing sites, all distributed in a rectangular grid.

4.2 Results

Figures 4a and 4b show the site amplification maps calculated using PGAs. For all the datasets, UPM has been compared with Kriging maps. When the number of observations (*N*) is low, the UPM has a smooth character with gradual transitions between the site amplification values as compared to the Kriging map. However, as the number of observations increases, the two maps start to become increasingly similar. If we focus on how the UPM changes with the increase in the number of observations, we observe that spatial variation starts appearing on the map and starts to converge as the number of observations (earthquake events) increases. To discuss this convergence quantitatively, Fig. 4c shows a plot of ΔD_{KL} with the *N*, the number of observations. The ΔD_{KL} is calculated only for the available sites common to all the events. It is shown that as the number of observations increases, the ΔD_{KL} decreases and starts to approach the minimum zero value. From the viewpoint of information theory, it can be concluded that the data is approaching saturation. We can then manage the seismic network, e.g., the observation period, and rearrange the layout to resolve the map in an unclear area, based on UPM.

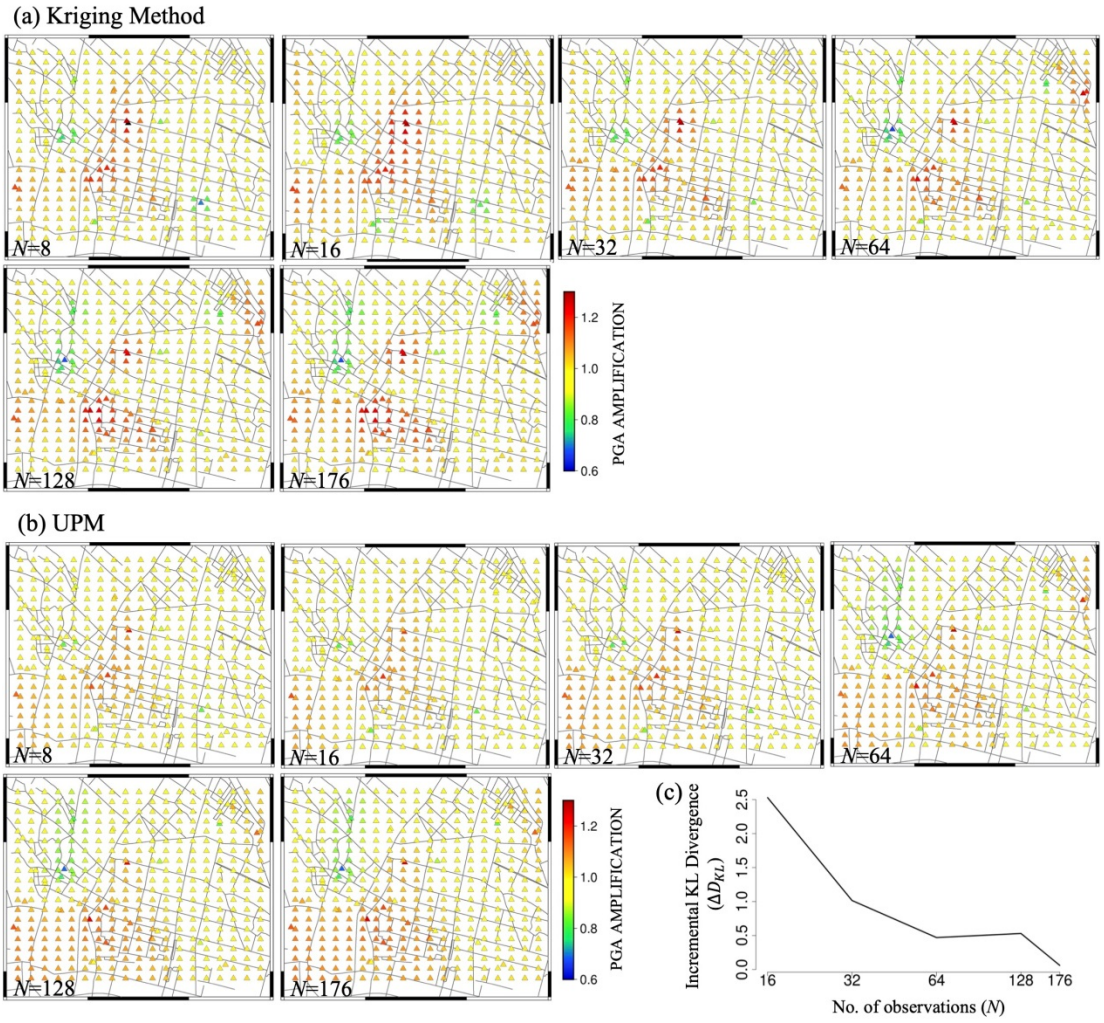

Fig. 4. (a) Evolution of Kriging maps of PGA amplifications in Furukawa district, Japan
(b) Evolution of UPM maps of PGA amplifications in Furukawa district, Japan
(c) Plot of ΔD_{KL} vs *N* for the UPM maps of PGA amplifications

Figures 5a and 5b show the site amplification maps calculated using PGVs. As before, both the UPM and Kriging maps have been prepared for the datasets. The first glance shows the PGV plots are smoother in comparison to the PGA plots. As observed in the case of the PGA plots, in this case too the UPM plots start to converge with the increasing number of observations. Figure 5c confirms that the data is approaching convergence from the viewpoint of information theory.

Thus, both the site amplification plots conclude that the mapping in Furukawa is approaching data saturation and based on the viewpoint of information theory, the current operation may be terminated. The seismometers may be rearranged to resolve the unclear areas.

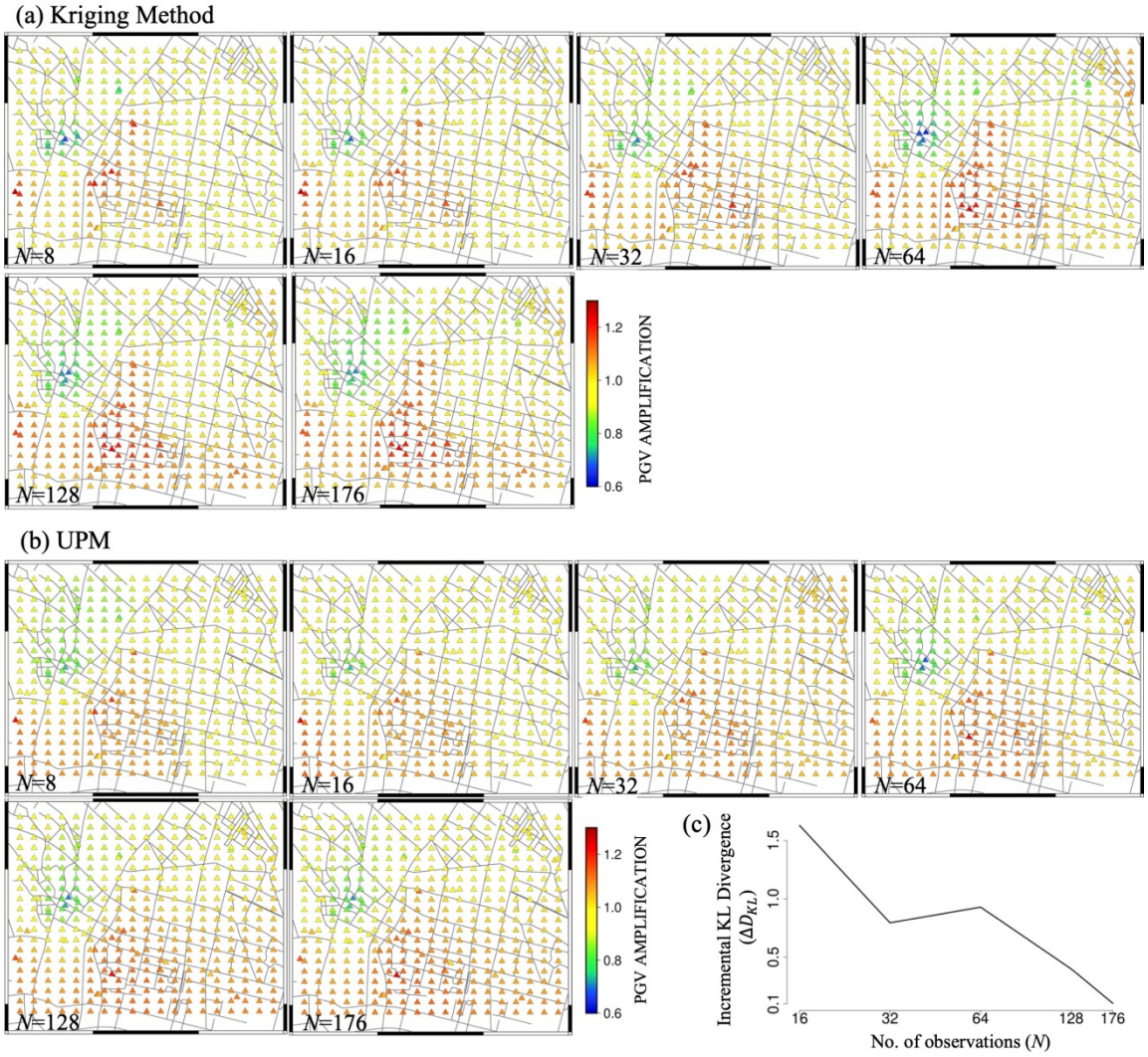

Fig. 5. (a) Evolution of Kriging maps of PGV amplifications in Furukawa district, Japan
(b) Evolution of UPM maps of PGV amplifications in Furukawa district, Japan
(c) Plot of ΔD_{KL} vs N for the UPM maps of PGV amplifications

5. DISCUSSION

It is evident from the cases discussed in Sections 3 and 4, that the optimum number of data which is deemed sufficient to extract useful information depends on the available dataset. In the case of numerical experiments, data saturation is attained after 512 observations have been collected. However, in the case of the seismic array in Furukawa, Japan, 176 observations seem to be sufficient to understand the population statistics.

The reason for this difference might be explained based on the record to record variability present at the sites. In the numerical experiment, the peak of true record to record variability was high (σ_j=3) in the boundary zone. Thus, more data is necessary to accurately estimate the mean and the record to record variability (standard deviation) in the boundary zone. However, although we will never know the true value of the population statistics for the case study area in Furukawa, Japan, the maximum estimated record to record variability recorded at any site was much lower and hence, lesser data was required to extract the desired information. Thus, the optimum number of data will vary from case to case and is most likely to be affected by the presence of high uncertainty zones.

For the case study area in Furukawa, Japan, we used 400 missing sites to create the PGA and PGV site amplification maps from 31 measurement sites, and the convergence process was quantified based on the maps obtained. However, the convergence process would not change if the grid size was any different. This is because the convergence is quantified considering the measurement sites only. Increase or decrease of grids using only missing sites will have no effect on the quantification of the convergence process.

Also, the convergence processes for PGA and PGV site amplification maps are not identical. One reason could be that the PGA and PGV processes are not the same. The spatial distribution patterns in Fig. 4 and Fig. 5 are clearly different. This means that the spatial datasets of PGA and PGV are different. Another reason could be the difference in the information gain process for the two processes. Unless two processes have the same incremental information gain and the same record to record variabilities at all locations, it will be rare for them to have the same convergence process based on information theory.

6. CONCLUSION

The availability of data has increased over the recent decades. However, we cannot ascertain whether the amount of available data is sufficient, and we have no guidelines to plot the maps based on the available data consistent with the data accumulation. In this study, we addressed these issues in terms of data visualization techniques.
We adopted UPM, which is a recently introduced mapping model that projects data uncertainties onto the map resolutions and hence, is more reliable statistically. As a measure of data saturation, we define a parameter ΔD_{KL}, based on information theory, which quantifies information gain as maps are updated with new data over time. Data saturation occurs when ΔD_{KL} approaches zero, which means that no more spatial information is being added to the maps and we can stop updating them.

The concept of visualizing data saturation was introduced as a numerical experiment using a simple model of wave amplification in and around an alluvial valley. The boundaries between the rock site and alluvial valley are high uncertainty zones. The results show that as we increase the number of observations, UPM starts converging with the Kriging map. This is a significant finding as it shows that UPM yields reliable results as compared to conventional mapping when less information is available and can be used to hint at data saturation as the number of observations increases. Measuring the change in ΔD_{KL} with the increasing number of observations, we found data saturation occurs after 512 observations have been collected.

The concept was then applied to a case study area in the Furukawa district of Japan where earthquake

data has been collected for over 7 years from 31 seismometers in a dense seismic array. Convergence in site amplification maps generated over different observation periods conclude that the mapping in Furukawa district is approaching data saturation and from the viewpoint of information theory, the current operation may be terminated and the seismometers may be rearranged to resolve mapping in the unclear areas.

Acknowledgement

We are grateful to Prof. Sumio Sawada from the Disaster Prevention and Research Institute, Kyoto University, Japan for his valuable comments and insights on the manuscript. We would also like to thank the two anonymous reviewers whose comments helped improve and clarify the manuscript. This work was supported by KAKENHI, Japan Society for the Promotion of Science (19H02224).

Computer Code Availability

Software Required: R, WinBUGS
Code Access: https://github.com/anirban1990/Visualising_DataSaturation_inUPM

Data availability

The seismic array data from Furukawa, Japan used in the application (Section 4) can be downloaded from: http://sn.catfish.dpri.kyoto-u.ac.jp/event_list/index.html. In total, there are 37 seismometers installed in the area. However, in this study, 31 seismometers that are in the significantly damaged area were utilized. The seismometers not considered in this study are F15, F21, F30, F32, F36 and F37.

REFERENCES

Banerjee, S., B.P. Carlin and A.E. Gelfand, 2014. Hierarchical modeling and analysis for spatial data. CRC Press.

Brodlie K., R.A. Osorio and A. Lopes, 2012. A Review of Uncertainty in Data Visualization. In Expanding the Frontiers of Visual Analytics and Visualization. Springer, London

Chakraborty, A. and H. Goto, 2018. A Bayesian model reflecting uncertainties on map resolutions with application to the study of site response variation. Geophysical Journal International, 214(3), 2264-2276.

Chaudhuri, S., R. Motwani and V. Narasayya, 1998. Random sampling for histogram construction: How much is enough? ACM SIGMOD Record, 27(2), 436-447.

Fusch, P. I. and L.R. Ness, 2015. Are we there yet? Data saturation in qualitative research. The Qualitative Report, 20(9), 1408-1416

Gelman, A., J. Hwang and A. Vehtari, 2014. Understanding predictive information criteria for Bayesian models. Statistics and Computing, 24(6), 997-1016.

Gilks, W. R., S. Richardson and D. Spiegelhalter, 1995. Markov chain Monte Carlo in practice. Chapman and Hall/CRC.

Goto, H. and J. Bielak, 2008. Galerkin boundary integral equation method for spontaneous rupture propagation problems: SH-case. Geophysical Journal International, 172(3), 1083-1103.

Goto, H. and H. Morikawa, 2012. Ground motion characteristics during the 2011 off the Pacific coast of Tohoku earthquake. Soils and Foundations, 52(5), 769-779.

Goto, H., H. Morikawa, M. Inatani, Y. Ogura, S. Tokue, X.R. Zhang, M. Iwasaki, S. Sawada and A. Zerva, 2012. Very dense seismic array observations in Furukawa district, Japan. Seismological Research Letters, 83(5), 765-774.

Guest, G., A. Bunce and L. Johnson, 2006. How many interviews are enough? An experiment with data saturation and variability. Field methods, 18(1), 59-82.

Harrower, M., 2003. Representing uncertainty: Does it help people make better decisions. In UCGIS Workshop: Geospatial Visualization and Knowledge Discovery Workshop (pp. 18-20).

Hughes, J. P. and D.P. Lettenmaier, 1981. Data requirements for kriging: estimation and network design. Water Resources Research, 17(6), 1641-1650.

James, B. R. and S.M. Gorelick, 1994. When enough is enough: The worth of monitoring data in aquifer remediation design. Water Resources Research, 30(12), 3499-3513.

Kawase, H., 1996. The cause of the damage belt in Kobe: "the basin-edge effect," constructive interference of the direct S-wave with the basin-induced diffracted/Rayleigh waves. Seismological Research Letters, 67(5), 25–34.

Kullback, S. and R.A Leibler, 1951. On information and sufficiency. The Annals of Mathematical Statistics, 22(1), 79-86.

Lee, J. G. and M. Kang, 2015. Geospatial big data: challenges and opportunities. Big Data Research, 2(2), 74-81.

Matheron, G., 1963. Principles of geostatistics. Economic Geology, 58(8), 1246-1266.

Stone, M., 1974. Cross-validation and multinomial prediction. Biometrika, 61(3), 509-515.

Trifunac, M. D., 1971. Surface motion of a semi-cylindrical alluvial valley for incident plane SH waves. Bulletin of the Seismological Society of America, 61(6), 1755-1770.

Wang, J. F., A. Stein, B.B. Gao and Y. Ge, 2012. A review of spatial sampling. Spatial Statistics, 2, 1-14.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11(Dec), 3571-3594.