17th World Conference on Earthquake Engineering, 17WCEE
Sendai, Japan - September 13th to 18th 2020

VISUALIZING DATA SATURATION IN GEOSPATIAL MAPPING WITH APPLICATION TO EARTHQUAKE ENGINEERING

A. Chakraborty[(1)], H. Goto[(2)]

[(1)] *Graduate Student, Graduate School of Engineering, Kyoto University, anirban@catfish.dpri.kyoto-u.ac.jp*
[(2)] *Associate Professor, Disaster Prevention Research Institute, Kyoto University, goto@catfish.dpri.kyoto-u.ac.jp*

Abstract

In earthquake engineering, mapping is frequently done to visualize the spatial variation of seismic site amplification. In conventional mapping, observations recorded over many seismic events at a site are reduced to a single averaged value. Thus, although site amplification at two sites might look visually different, statistical significance of this difference is directly ungraspable without any information on the data uncertainty. Uncertainty Projected Mapping (UPM) adds statistical significance to mapping by projecting data uncertainty on map resolutions. Here, we introduce the principal concept of UPM and discuss its data dependency property. UPM approaches conventional mapping as data increase and we use it to quantify data saturation in geospatial mapping. A parameter is proposed that measures the incremental information gain as new data is added to mapping. Data saturation is reached when the proposed parameter approaches zero. The concept is applied to a seismic array in Furukawa district of Japan where seismic data is collected over 7 years from 31 seismometers. Convergence in site amplification maps generated over different observation periods conclude that the mapping in Furukawa district is approaching data saturation and from the view point of information theory, the current operation of seismic monitoring may be terminated.

Keywords: data saturation; geospatial statistics; site amplification; uncertainty; uncertainty projected mapping

1. Introduction

In earthquake engineering, mapping is frequently done to visualize the spatial distribution of many different variables. Immediately after an earthquake, USGS releases ShakeMaps which provide near-real-time spatial distribution maps of ground motion and shaking intensity. Long term earthquake records are utilized to generate hazard maps and highlight areas vulnerable to significant earthquake damage. The knowledge of spatial distribution of shear wave velocity is crucial in identifying sites susceptible to strong ground shaking. Many variants of these geospatial maps are used by several organizations for post-earthquake response, preparedness exercise and disaster planning, etc.

However, the resolution of these spatial maps is not always reliable at local scales. Fig. 1 shows a local scale map of spatial distribution of site amplification factor (mapped variable) in an area of Japan [1]. Let us focus at the situation at A, where blue and red colored sites, representing extreme site amplification factors, are situated right next to each other. How reliable is this situation? Is it possible to explain if this situation belongs to case 1 or case 2? If it is case 1, where the difference in neighboring values is statistically significant (non-overlapping data distributions), situation at A is reliable. However, if it is case 2, where the difference in neighboring values is not statistically significant (overlapping data distributions), the color separation at A is not reliable. Unfortunately, the conventional maps cannot distinguish between the cases 1 and 2, as the information of data variation (uncertainty) is not included in the mapping process. Situations like this is not uncommon in spatial maps. The inability of conventional maps to statistically signify the difference in mapped values, raises a question on its use for reliable decision making process.

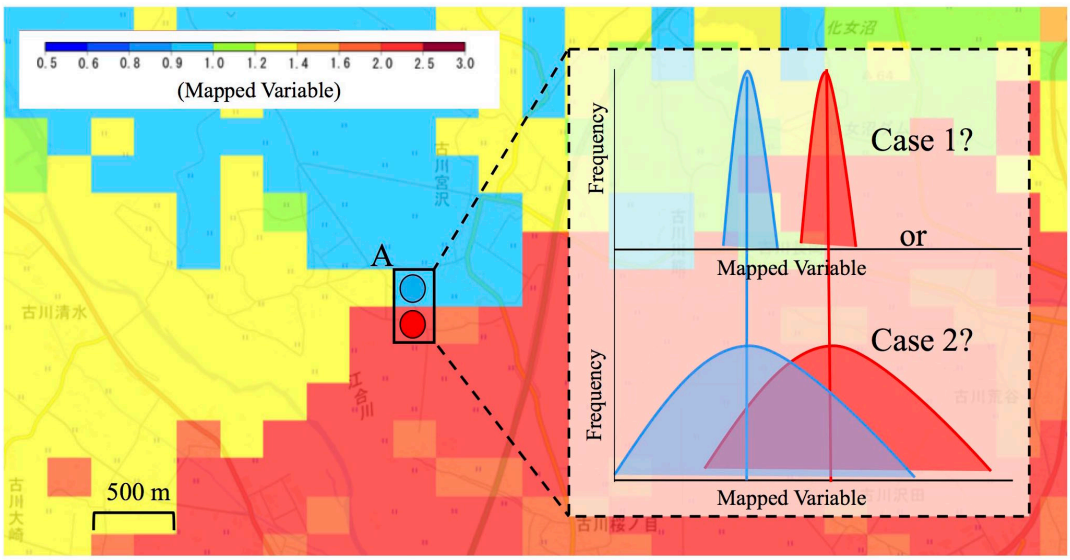

Fig. 1- A local scale spatial distribution map [1]

To handle the reliability issues with spatial distribution maps, Chakraborty and Goto [2] proposed Uncertainty Projected Mapping (UPM), where in addition to the mean value, uncertainty information is projected in the map resolutions in such a way that map resolutions in high uncertainty zones are colored smoothly. In UPM, a smoothing effect is introduced and map resolutions are reduced in high uncertainty zones where the difference in neighboring values are not statistically significant. However, the map resolution remains high in low uncertainty zones where the difference in neighboring values is statistically significant. Thus, if Fig. 1 was replotted with UPM, and the data distribution belonged to case 2, instead of two distinct colors (blue and red) at A, there will be only one color at A. UPM map is a better representation of the available data and can help in reliable decision making process.

Another important issue with mapping is the issue of data sufficiency. Data has increased significantly in recent times. However, it is usually not clear if the amount of data is enough to extract the desired information. Continuing data collection adds to the computational cost as the new data being processed might be redundant. An interesting data dependent property of UPM maps, which will be discussed throughout this paper, offers a solution to this issue. UPM maps evolve with addition of new data, and after a while they start converging. In this study, we quantify the convergence process by measuring the incremental information gain as maps are updated with new data. Point of data sufficiency (data saturation) is assumed when no more incremental information gain happens even after adding new data to a map.

The purpose of this paper is twofold. Firstly, to introduce the basic concept of UPM [2] and then to discuss the data dependency property of UPM which plays an important role in visualizing the data saturation in mapping. Although UPM and visualizing data saturation in UPM maps are applicable to any spatial variable, in this paper, we focus on site amplification as the spatial variable. In the numerical experiment, we study the site amplification variation in a one-dimensional alluvial basin. And as a case study, we study the site amplification in Furukawa district of Japan using long term data from a dense seismic array being operated there.

2. Uncertainty Projected Mapping(UPM)

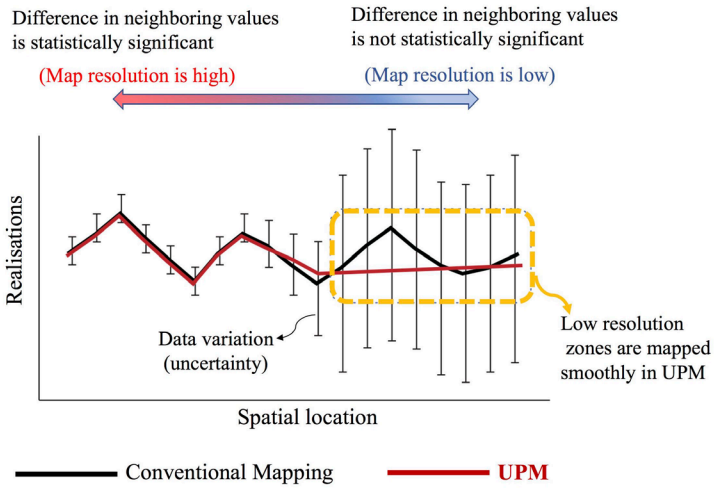

Fig. 2- Basic concept of Uncertainty Projected Mapping(UPM)

In this section, we introduce the basic concept of UPM. In Fig. 2, conventional mapping varies as a triangular wave in one-dimensional space. In this paper, for conventional mapping, we use Kriging [3], a popular tool of spatial interpolation. The uncertainty at site j (standard deviation, σ_j) increases from left to right. As discussed in section 1, the goal of UPM is to reflect the uncertainty information into the map resolutions and reduce the map resolution in zones of high uncertainty. Thus, in the left zone, where the data variation is low, UPM follows the conventional mapping and maintains a high map resolution. However, in the right zone, where the data variation is high, a smoothness is introduced and the map resolutions decrease.

To obtain the smoothing effect in zones of high uncertainty, UPM considers two uncertainties: record to record variability at a site j (σ_j) and site to site variability in the neighborhood of j (s_j) . These two uncertainties are related such that

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

where c is a constant. In UPM, s_j helps make the map resolutions smooth where σ_j is high. A low s_j value means low variability around j and hence, a smooth resolution. However, a high s_j means a high

variability around j and thus, a rough resolution. At zones of high σ_j we impose a low s_j and make the resolutions smooth. At zones of low σ_j, we impose a high s_j and make the resolutions rough. The constant c is unique to a model setting and the optimum value of the constant c is based on model evaluation.

UPM is modelled as a Bayesian hierarchical model [4]. The unknown parameters μ, σ and s are assigned a prior distribution and estimated based on a posterior probability distribution using Markov Chain Monte Carlo algorithms. However, many different UPM can be generated based on different c values. Model evaluation is done using widely adopted information criterion (WAIC) [5]. 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 points, if necessary) so that every site almost has the same number of sites in the neighborhood.

3. Methodology

3.1 Update UPM maps at multiple stages of data accumulation

The first step in visualizing data saturation is to create a series of UPM Maps at different stages of data accumulation. UPM maps evolve with the increasing data as the estimation of mean (μ_j) and record to record variability (σ_j) depends on the amount of data. As we will see in section 4 and section 5, the UPM maps converge with conventional mapping as the data increase. Quantifying this convergence process will help us find at which stage data saturation occurs.

3.2 Measuring the incremental information gain as the UPM maps evolve

To quantify this convergence in UPM maps, we use a parameter based on Kullback-Leibler (KL) Divergence [6]. KL Divergence measures how different two probabilistic distributions are. 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

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

where $\Delta D_{KL_{[N+\Delta N]}}$ is the D_{KL} between the probability distribution $P_{[N]}$ at N observation data case and the probability distribution $P_{[N+\Delta N]}$ at $N + \Delta N$ observation data case summed over the j sites.

The parameter ΔD_{KL} measures the incremental information gain as the maps are updated with more and more data in time. Data saturation happens 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.

4. Numerical Experiment

4.1 Data

As numerical experiment, we model the spatial variation of wave amplification (hypothetical) in a one-dimensional alluvial basin. Alluvial deposits can significantly affect the amplitudes of the incident seismic waves in a basin [7]. It is well known (low uncertainty) that the incident seismic waves get highly amplified at the center of an alluvial basin. However, at the basin edge (boundary between the rock site and alluvial

basin) there is a high uncertainty regarding the wave amplification properties because the incident angle and frequency contents are well affected. In Fig. 3, randomly generated samples at each site are wave amplification values for different incident seismic waves. The samples at each site follow a lognormal distribution. The well accepted trend of wave amplification variation is captured by the trend of mean (μ) variation, which increases from the edge to the center of the basin. The uncertainty knowledge is captured through the record to record variability (σ), which is high at the basin edge, and low at all other locations.

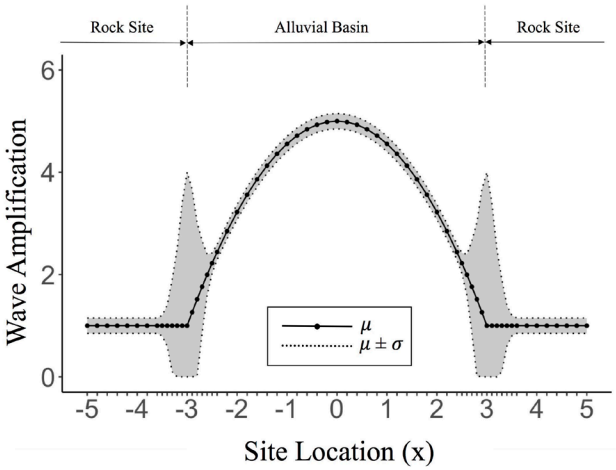

Fig. 3- Numerical experiment: A hypothetical model of wave amplification in an alluvial basin

4.2 Results

Eight UPM maps with 8, 16, 32, 64, 128, 256, 512 and 1024 samples (gray dots) are created (Fig. 4a). Each succeeding dataset includes the preceding dataset with lower earthquake events.

It is observed that when the number of observations (*N*) is low, UPM maps show a smooth transition at the highly uncertain basin edge, where the Kriging map is very rough and fluctuating. This smoothness is introduced by Eq. (1). However, as the number of observations (*N*) increase, UPM maps starts to converge with Kriging maps. This change in the characteristics of UPM with the increase in *N* has a significance in understanding the population.

When *N* is low, there is less information for modelling and so, the estimated model parameters are quite unstable. The estimates are erroneous in high uncertainty zones. In such a situation, the smooth UPM maps in the highly uncertain basin edge is a better representative of the physical process than the erroneous Kriging maps.

When *N* is high, there is more information for modelling and so, the estimated model parameters are stable. Due to increased data, error is also reduced in high uncertainty zones. It is very interesting to observe that the UPM maps now converge with the Kriging maps. Thus, UPM yields reliable results as compared to Kriging when less information is available and can be used to hint at data saturation as the number of observation increases.

The change in incremental KL divergence (ΔD_{KL}) with respect to *N* is shown in Fig. 4b. Sites located at the edges are not included in the calculation of ΔD_{KL}. It is observed that ΔD_{KL} starts to converge as *N* increases. This indicates that the UPM maps reaches convergence and the data set is sufficient to extract the population statistics.

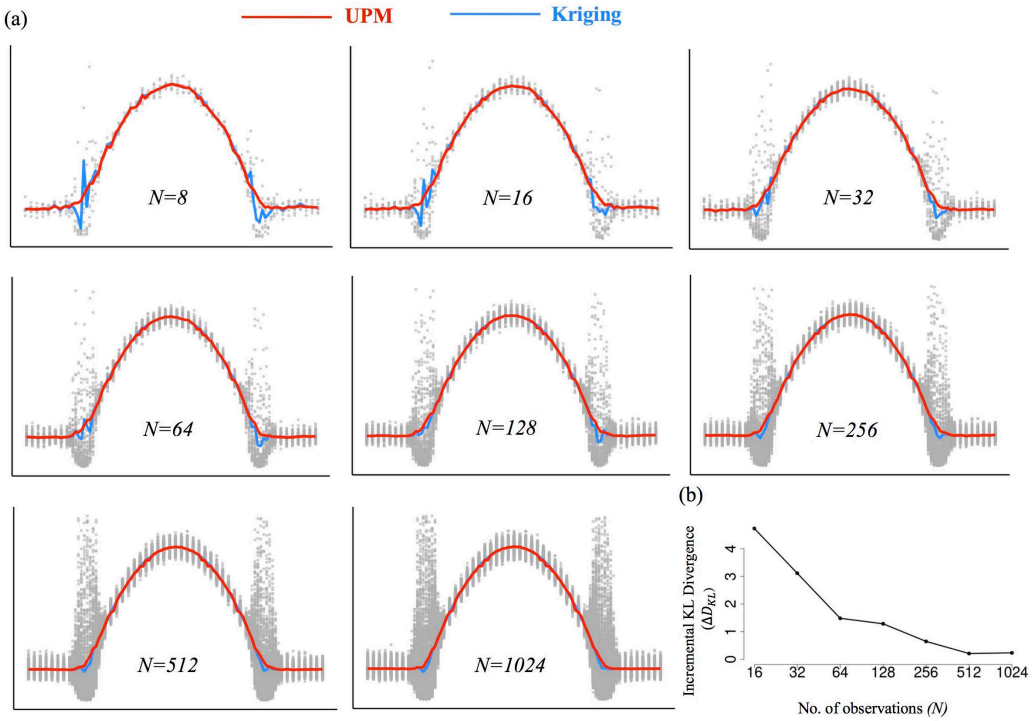

Fig. 4- (a) Evolution of UPM and Kriging maps for numerical experiment
(b) Plot of ΔD_{KL} vs N for the UPM maps

5. Case study: Site Amplification in Furukawa district, Japan

5.1 Data

In this case study, long term earthquake data from a dense seismic array is used to create a series of local scale site amplification maps. The dense seismic array is situated in Furukawa district of Japan [8]. During the 2011 off the Pacific coast of Tohoku Earthquake, downtown residential areas of Furukawa district incurred severe damages mainly due to site amplification. Significant spatial differences of ground motion were observed even in sub-kilometer scale and hence, since 29th October, 2011, a dense seismic array is being operated for in-depth study of the area.

Fig.5 shows the layout of the dense seismic array in Furukawa district. In total, there were 37 seismometers. However, we focus on the 31 seismometers situated in the significantly damaged downtown area. As for the observation data, we use 176 earthquake events collected over a period of 7 years (29th October, 2011 to 19th September, 2018). 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 threshold of amplitude or condition of source location. Also, each of the 31 seismometers didn't record all the 176 earthquake events. For studying the convergence process, 6 datasets were created using groups of 8,16,32,64,128 and 176 earthquake events. Each succeeding dataset includes the preceding dataset with lower earthquake events.

The mapped variable is a site amplification factor observed at site j during an earthquake event. It is defined as the logarithmic ratio of observed peak ground velocity (PGV) at site j to the spatial average calculated over all the available sites during an earthquake event. The PGV is calculated from the vector sum

of EW component and NS component of the earthquake record. To generate a UPM map of the site amplification factor, the dataset comprised of 431 sites with 31 measurement sites from the seismic network and 400 missing sites, all distributed in a rectangular grid.

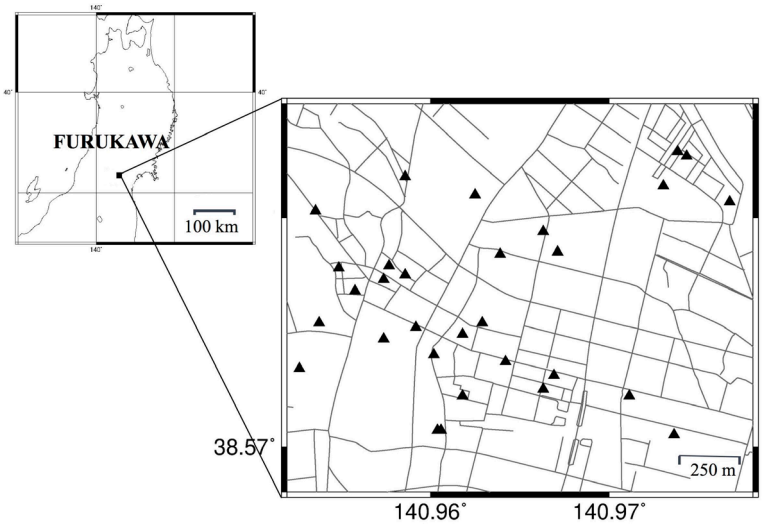

Fig. 5- Spatial distribution of seismometers (▲) in Furukawa district, Japan

5.2 Results

Fig. 6a and Fig. 6b show the site amplification maps calculated using PGVs. Kriging maps are created in addition to the UPM maps for comparison. It is observed that when the number of observations (N) is low, the UPM maps are smooth with gradual transitions between the site amplification values as compared to the Kriging maps. However, as the number of observations (N) increase, spatial variation starts appearing on the UPM maps and they start to converge. To discuss this convergence quantitatively, Fig. 6c shows a plot of ΔD_{KL} with N. The ΔD_{KL} is calculated only for the sites common to all the events. It is shown that as the number of observations increase, 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 the unclear area, based on UPM.

6. Discussion and Conclusion

In recent times, data is becoming more accessible. Visualization is becoming more important and geospatial maps are now taking a common place in many different fields. Many of these maps are used in decision making process. The conventional visualization techniques assume that the data is free of uncertainty and the resolutions are not always reliable. UPM maps which project uncertainty in map resolutions can lead to more reliable decision making and has application in a wide range of problems in earthquake engineering.

The numerical experiment and case study reveal 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 observation increases. It is also evident that the optimum number of data which is deemed enough to extract useful information depends on the available dataset. In the case study problem, data sufficiency is reached much earlier as compared to the numerical experiment. It is probably because optimum data necessary to accurately estimate the mean and the record to record variability is affected by the presence of high uncertainty zones.

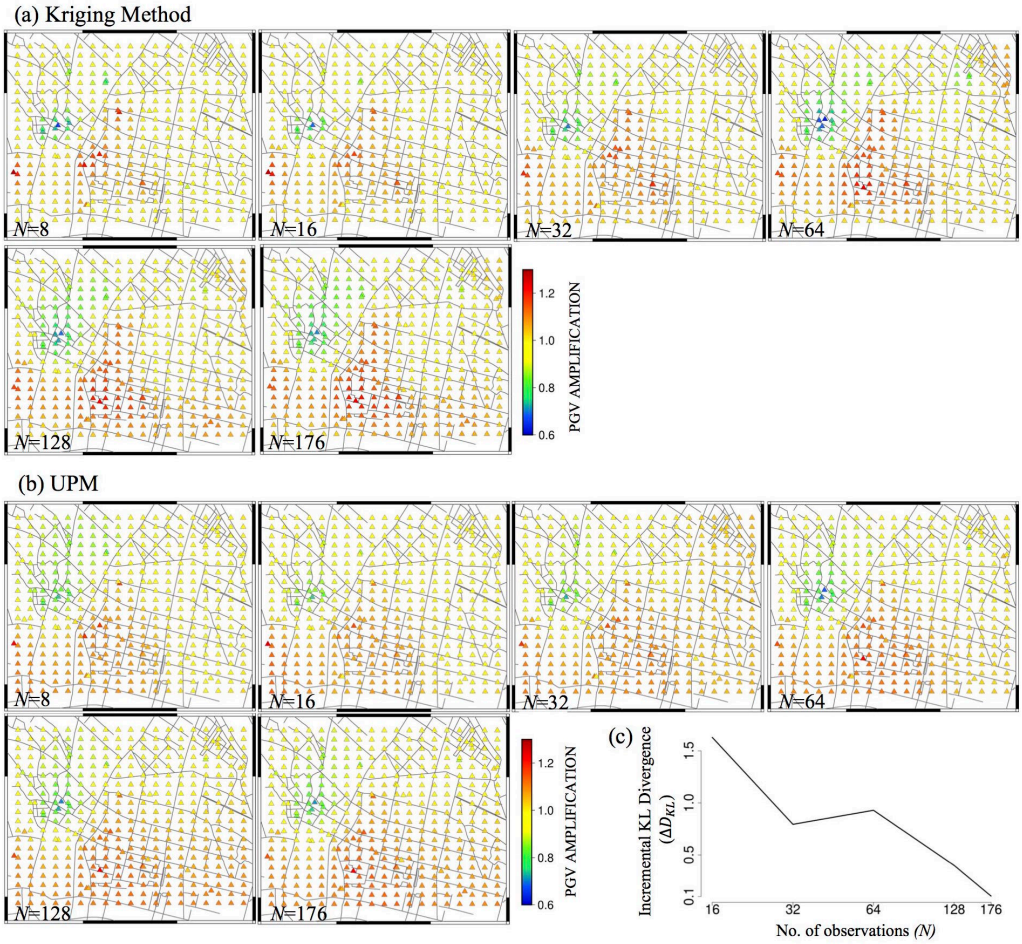

Fig. 6-(a) Evolution of Kriging maps of PGV amplifications in Furukawa district, Japan (b) Evolution of UPM maps of PGV amplifications in Furukawa district, Japan (c) Plot of in ΔD_{KL} vs N for the UPM maps of PGV amplifications

7. Acknowledgement

The authors are grateful to Prof. Sumio Sawada from Disaster Prevention and Research Institute, Kyoto University, Japan for his valuable comments and insights on the manuscript. This work was supported by KAKENHI, Japan Society for the Promotion of Science (19H02224)

8. References

[1] National Research Institute for Earth Science and Disaster Prevention (NIED) (2016): *Japan Seismic Hazard Information Station (J-SHIS)*. http://www.j-shis.bosai.go.jp/

[2] Chakraborty A, Goto H (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.

[3] Matheron G (1963): *Principles of geostatistics*. Economic geology, **58**(8), 1246-1266.

[4] Banerjee S, Carlin BP, Gelfand AE (2014): *Hierarchical modeling and analysis for spatial data*. CRC Press, 2nd edition.

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

[6] Kullback S, Leibler RA (1951): *On information and sufficiency*. The annals of mathematical statistics, **22**(1), 79-86.

[7] Trifunac MD (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.

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

