# Map Resolutions considering Data Uncertainty with Application to Seismic Microzonation

CHAKRABORTY ANIRBAN

2021

## ABSTRACT

In earthquake engineering, seismic microzonation is important to mitigate a future earthquake disaster. The spatial variation of site response is caused primarily due to varying amplification of the seismic wave by the ground surface layers. In practice, the spatial variation of site response is projected on maps. However, the resolution of these spatial maps is not always reliable at local scales. Conventional mapping techniques assume that the data is free of uncertainty and uses only the mean value at a site. The site responses between two sites with some difference in their average value are visually considered to be different. However, the statistical significance of this difference is directly ungraspable without any information on the data uncertainty.The inability of conventional maps to statistically signify the difference in mapped values, raises a question on its use for reliable decision-making process.

Many researchers believe that including uncertainty on maps could lead to better decisions. However, in literature, I hardly came across any research addressing this issue of projecting data uncertainty onto the map resolutions. This study has three objectives. The first objective is to propose a methodology that projects data uncertainty onto the map resolutions. The proposed mapping mapping methodology is named as "Uncertainty Projected Mapping" (UPM). The second objective is to investigate how the UPM map resolutions change as the number of observation data increase, and confirm if UPM really projects data uncertainty onto map resolutions. The third objective is to use the framework of UPM and update map resolutions of a conventional map using data uncertainty from local sources.

In the first study, a relation between the site-specific uncertainty and spatial uncertainty in the framework of a hierarchical Bayesian model is introduced. The idea is to make the spatial resolution low (or smooth) at zones of high data uncertainty. The proposed UPM methodology was validated with both numerical experiments and real data from a very dense seismic array. The UPM results were found to project the site-specific uncertainties in the map resolutions. The detailed visual (mapping) on significantly different observations expected between the sites with low data uncertainty and rough visual (mapping) on insignificant observations between the sites, was enhanced in the UPM maps. The UPM methodology could spatially interpolate and estimate values at the missing sites. The mapping results were compared with a conventional mapping technique called Kriging. It was observed that unlike the UPM values which are sensitive to the variation of data uncertainty, the Kriging values are not affected by the change in data uncertainties. Thus, the map resolutions determined by UPM has a certain degree of statistical significance and might be considered reliable.

In the second study, I investigate how the UPM map resolutions change as the number of observation data increase and confirm that UPM really projects data uncertainty onto map resolutions. It is observed that as more and more information become available, UPM starts approaching the conventional mapping. This characteristic hints at the strength of UPM when less information is available. I investigate this characteristic in detail and utilize it to propose a parameter to measure the change in map resolutions with increasing information, which was applied for quantification of data saturation in mapping spatial data. The results show that the optimum number of data which is deemed enough to extract useful information depends on available dataset. This study establishes the fact that when the number of observations is less, UPM is a more reliable representation of the data.

In the third study, I apply UPM to incorporate local soil boring data uncertainty and update the map resolutions of J-SHIS map of site amplification factor in Ibaraki-Takatsuki area of Osaka. The J-SHIS map of site amplification factor is based on values of average shear velocity in the upper 30 m depth of soil (AVs30) broadly assigned to engineering geomorphic units. There is a need to incorporate local (site-specific) information in order to increase the reliability of map resolutions at local scale. Using seismic ground response analysis, I calculate site amplification factor from the available soil boring data in individual meshes. Further, using the Bayesian framework of UPM, I updated the existing map resolutions of J-SHIS map. The updated J-SHIS map reflects the data information and highlights significant differences, when such a difference exists, and for situations with low data where information cannot be extracted, a low spatial resolution (smooth mapping) is introduced adding more reliability (in statistical terms) in comparison to original J-SHIS map.

The results in all the three studies primarily establish that that map resolutions determined by data uncertainty has a statistical significance and hence is a more reliable representation of the data at a site, unlike conventional mapping techniques which reduce the data to a single value at a site.

# TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

## 1.1 Background and motivation

In earthquake engineering, understanding the spatial variation of site response is important. It is caused primarily due to varying amplification of the seismic wave by the ground surface layers. Past observations show an extreme variability of the ground motion during an earthquake event. One of the first instances when such a variation in damage was recorded was during the 1985 Michoacán earthquake in Mexico, where very soft sediments caused long-period and long-duration strong ground motions and severe damage at a location 400 km away from the epicenter [**1,2,3**]. Also, during the 1989 Loma Prieta earthquake in United States, a strong correlation was identified between the damage and the local geology of surface sediments [**4,5**]. In Japan, a narrow damage band, where more than 30% of wooden residences collapsed, appeared in Kobe and the surrounding cities, during the 1995 Kobe earthquake [**6**]. However, an earthquake of similar magnitude in Tottori hardly caused any damage [**7,8**]. Serious damage was observed to be concentrated in Adapazari town in Turkey during the 1999 Kocaeri earthquake in Turkey [**9**]. The damage observed in Christchurch in New Zealand during the 2010-2011 Canterbury earthquake series was also related to the variation of local site conditions [**10,11**]. During the 2011 off the Pacific coast of Tohoku Earthquake in Japan, Furukawa district in Osaki city incurred serious damage concentrated in a small area of 0.5 ×1.0 $km^2$ [**12**]. The Kathmandu valley in Nepal experienced significant seismic motions during the 2015 Gorkha earthquake [**13**]. The ground motions were exceptionally high due to the amplification caused by soil conditions [**14, 15**], which lead to widespread significant structural and geotechnical damage [**13, 16**]. The spatial difference in site amplification of ground surface caused heavy damage limited to Mashiki town of Kumamoto, during the 2016 Kumamoto earthquake [**17,18**].

Seismic hazard maps, based on models and observations, are an appealing tool for evaluation of earthquake hazard and risk. One of the earliest reasons to push for the qualitative evaluation of hazard was the 1933 Long Beach earthquake in California, USA which served as a wake-up call to engineers [**19, 20**]. A tradition was established in which experience gained in significant earthquakes is incorporated into subsequent updates to the building codes. Through the 1940s and 1950s, seismic design provisions in building codes tended to be based on qualitative evaluations of hazard. Later,

quantitative seismic hazard maps based on probabilistic analysis were introduced. In Canada, for example, a major watershed for seismic design philosophy came in 1970 with the inclusion of the first national probabilistic seismic hazard map based on the work of Milne and Davenport [**21**]. Since the 1970s, seismic hazard maps have been developed for building code applications based on a probabilistic approach. Around the same time that Milne and Davenport [**21**] were developing their seismic hazard maps of Canada, Cornell [**22**] was developing a somewhat different methodology, which was coded into a FORTRAN algorithm by McGuire [**23**]. Today, almost all earthquake prone countries and regions of the world have a probabilistic seismic hazard map, although the level of knowledge and details may vary considerably based on the available data [**24,25, 26, 27, 28, 29**]. As national seismic hazard maps started to become commonplace, a new attempt started to bring together the knowledge spread over different regions into a unified platform. One of the earliest global seismic hazard maps was generated by the Global Seismic Hazard Assessment Program (GSHAP) [**30**]. Another initiative, the global earthquake model (GEM) generates the Global Earthquake Risk Map [**31,32,33**].

All these hazard maps represent some sort of a spatial distribution of a variable. For example, the Global Earthquake Hazard Map in GEM depicts the geographic distribution of the Peak Ground Acceleration (PGA) with a 10% probability of being exceeded in 50 years, computed for reference rock conditions (shear wave velocity of 760-800 m/s). Similarly, ShakeMap, which is a product of the USGS Earthquake Hazards Program in conjunction with the regional seismic networks, provides near-real-time maps of ground motion and shaking intensity following significant earthquakes [**34**]. ShakeMaps are used by federal, state, and local organizations, both public and private, for post-earthquake response and recovery, public and scientific information, as well as for preparedness exercises and disaster planning. In Japan, the National Seismic Hazard Maps are prepared by the Headquarters for Earthquake Research Promotion (HERP). These maps are basically a spatial distribution of estimated strong motions from future earthquakes. The seismic hazard information has been made public on the Japan Seismic Hazard Information Station (J-SHIS) portal [**26**].

In seismic hazard assessment, the effects of local soil deposits on the seismic hazard are often treated with less rigor than this critical aspect deserves [**35**]. The study of the site response effects can be performed with varying levels of sophistication [**36**]. The most common approach uses soil classification [**37**] which is based either on the shear wave velocity in the uppermost 30 m of soil (Vs30) [**38**] or proxy information, for example, local topography or geo-lithology [**39,40**]. Seismic microzonation is important as it improves the seismic hazard assessment to be usable at local scales and with a higher resolution.

In order to establish the seismic microzonation at local scale, it is also important to pay attention to the map resolution to understand if the color difference is plausible or not. There is an issue of reliability with the spatial resolutions at local scales. The conventional mapping techniques only consider the fact that observations vary in space. However, the observations vary not only in space but also with each event. In one example, **Fig. 1.1** shows how the spatial patterns of peak ground acceleration (PGA) of

ground motion recorded during two different earthquake events, are not the same [**41**]. This variation among events is also referred to as the uncertainty of the observations at a site. Although there is a difference in the definition of uncertainty and data variability [**42**], for all practical purposes, in this study we include both under the heading of data uncertainty. In conventional mapping, the observations recorded over events at a site are generally reduced to a single averaged value. The site responses between two sites with some difference in their average value are visually considered to be different. However, the statistical significance of this difference is directly ungraspable without any information on the data uncertainty.

This issue in reliability at local scale is introduced using the J-SHIS map of site amplification factor, which is based primarily on the J-SHIS map of engineering geomorphic classification. The engineering geomorphic classification map offers the geomorphic classification with a spatial resolution of 250 m in whole of Japan [**43**]. Using the engineering geomorphic classification map, J-SHIS map of average shear velocity in the upper 30 m depth (AVs30) is calculated [**44**]. Finally, the J-SHIS map of site amplification factor is prepared based on the AVs30 values [**45**]. The amplification factor means amplified ratio calculated from the engineering bedrock (Vs=400 m/s) up to the ground surface.

**Fig. 1.2** shows the J-SHIS map of site amplification factor at a local scale [**46**]. Let us focus at the situation at A, where blue and red colored mesh, representing extreme site amplification factors, are situated right next to each other. How reliable is this spatial resolution? Is it possible to explain if this spatial resolution 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), the existing spatial resolution 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, and a smooth or low spatial resolution might better explain the situation. Unfortunately, the conventional maps cannot distinguish between the cases 1 and 2, as the information of data uncertainty is not included in the mapping process. Situations like this are 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.

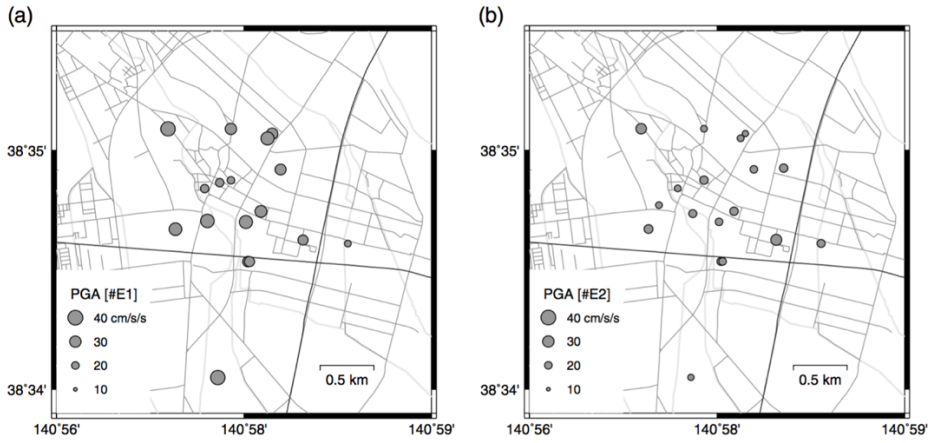



**Figure 1.1** Uncertainty in spatial distribution of PGA during two earthquake events [**41**]

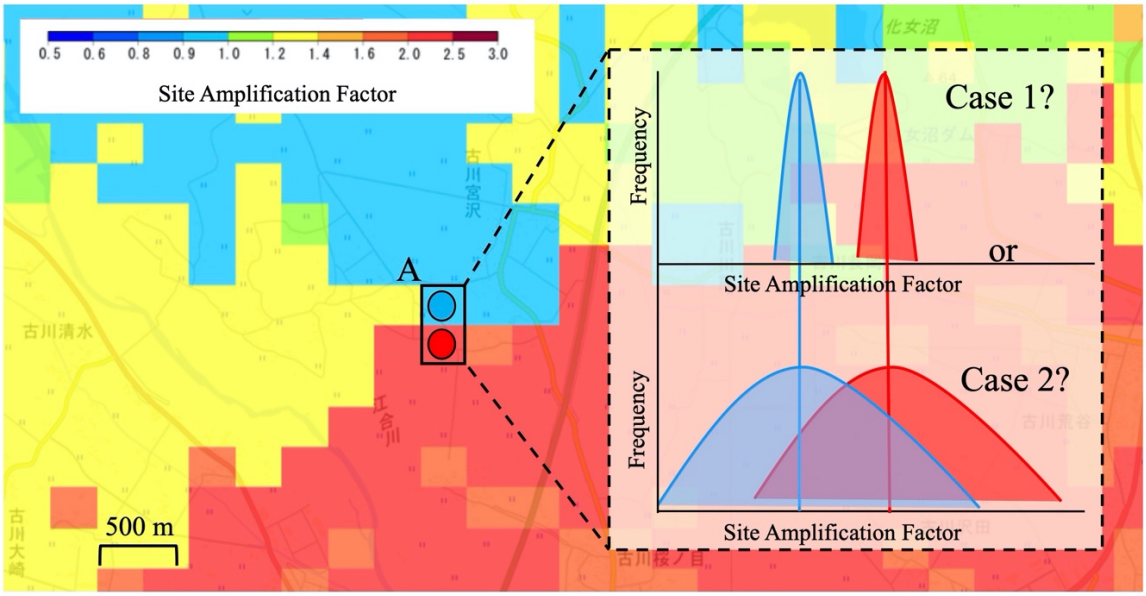



**Figure 1.2** Limitations of conventional map in validating map resolutions [**46**]

## 1.2 Scope of research

Many researchers believe that displaying uncertainty on maps could lead to better decisions [**47**]. Conventional visualization techniques assume that the data is free of uncertainty and uses only the mean value at a site [**48**]. However, in literature, there is hardly any paper addressing this issue of incorporating uncertainty in the map resolutions. Although there are some works on the uncertainty in seismicity, but they are quite different from the problem I want to address. Kuehn and Scherbaum [**49**] estimated a partially non-ergodic ground-motion prediction equation using a hierarchical model that

accounts for regional differences. Ordaz and Arroyo [**50**] discussed the issue of correct estimation of uncertainties in probabilistic seismic hazard analysis. They examined to estimate the uncertainty at the site but did not discuss their projection onto the map. De Risi et al. [**51**] stressed on the need for a better mapping process to extract more information from limited data. Rodriguez-Marek et al. [**52**] proposed a methodology to capture epistemic uncertainty in site response using the conventional logic tree approach for seismic hazard assessment. There is a scope of research in a methodology that determines map resolutions based on data uncertainty.

## 1.3 Objective of this study

This study has three objectives which are listed as follows:

1. The first objective of this study is to propose a methodology that reflects data uncertainty onto map resolutions. The proposed methodology uses a hierarchical Bayesian framework [**53**]. The basic concept of the proposed methodology is explained in **Fig. 1.3**. Both panels (**Figs 1.3a** and **b**) show the hypothetical spatial data set with the same mean patterns but with different data uncertainties. In conventional mapping, the map only follows the averaged value at the sites and the spatial resolution is independent of the data uncertainty. I propose a mapping that is sensitive to the data uncertainty. In **Fig. 1.3(a)** where the data uncertainty is low, the proposed mapping should follow conventional mapping (high spatial resolution) but as the uncertainty increases in **Fig. 1.3(b)**, the proposed mapping should no longer follow the conventional mapping but become smooth (low spatial resolution). The map reflects the statistical significance based on the data uncertainty and so I named the proposed mapping methodology as "Uncertainty Projected Mapping" (UPM).
2. The second objective of this study is to investigate how the UPM map resolutions change as the number of observation data increase, and confirm if UPM really projects data uncertainty onto map resolutions. As the number of observation data increase, the data uncertainty decreases and the reliability in the estimation of the mean increases. And if UPM projects data uncertainty onto map resolutions, this change in data uncertainty with increasing observations must be reflected in the spatial resolutions.
3. The third objective of this study is to use the framework of UPM and update map resolutions of a conventional map using data uncertainty from local sources. As an example of conventional map, I use the J-SHIS map of site amplification factor, and as local source of data uncertainty, I use soil boring data.

## 1.4 Outline of thesis

This thesis consists of six chapters (see **Fig. 1.4**). Chapter 2 introduces the mathematical background of Bayesian hierarchical modeling. Chapter 3 describes the proposed methodology of UPM in detail. Chapter 4 investigates how the UPM map resolutions change with increasing number of observations. Chapter 5 explains the use of UPM framework to update map resolutions of a conventional map. Chapter 6 summarizes the important points of this study.

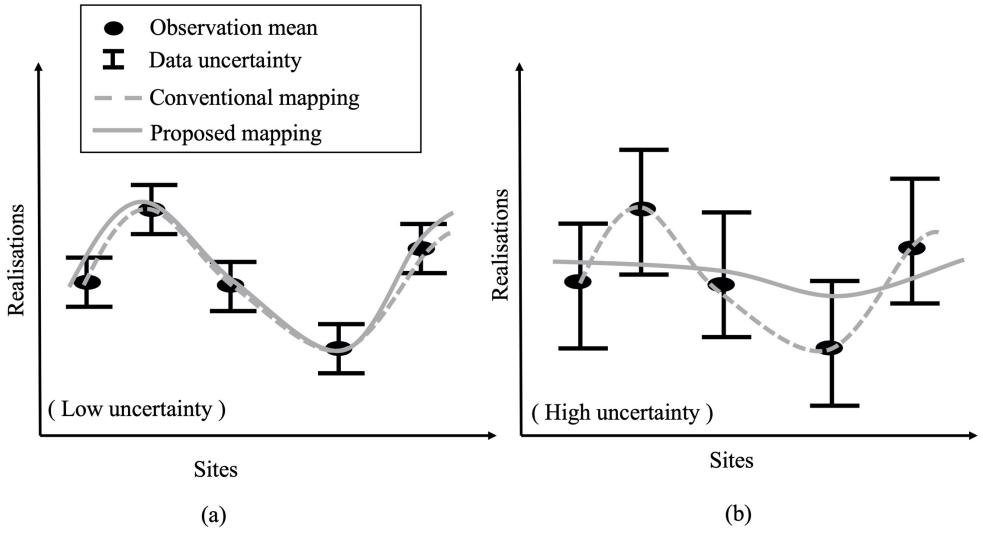



**Figure 1.3** The basic concept of the proposed UPM methodology

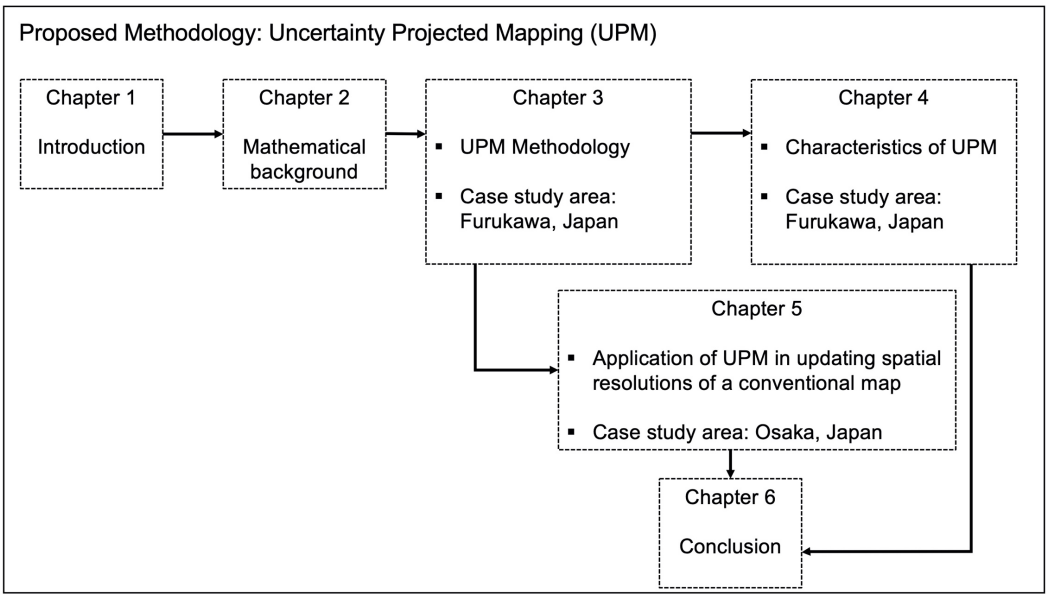



**Figure 1.4** Flow diagram of this study

## References


1. Campillo, M., Gariel, J. C., Aki, K., & Sanchez-Sesma, F. J. (1989). Destructive strong ground motion in Mexico City: Source, path, and site effects during great 1985 Michoacán earthquake. *Bulletin of the Seismological Society of America*, *79*(6), 1718-1735.
2. Singh, S. K., Lermo, J., Dominguez, T., Ordaz, M., Espinosa, J. M., Mena, E., & Quaas, R. (1988). The Mexico earthquake of September 19, 1985—a study of amplification of seismic waves in the valley of Mexico with respect to a hill zone site. *Earthquake spectra*, *4*(4), 653-673.
3. Singh, S. K., Mena, E. A., & Castro, R. (1988). Some aspects of source characteristics of the 19 September 1985 Michoacan earthquake and ground motion amplification in and near Mexico City from strong motion data. *Bulletin of the Seismological Society of America*, *78*(2), 451-477.
4. Chin, B. H., & Aki, K. (1991). Simultaneous study of the source, path, and site effects on strong ground motion during the 1989 Loma Prieta earthquake: a preliminary result on pervasive nonlinear site effects. *Bulletin of the Seismological Society of America*, *81*(5), 1859-1884.
5. Bonamassa, O., & Vidale, J. E. (1991). Directional site resonances observed from aftershocks of the 18 October 1989 Loma Prieta earthquake. *Bulletin of the Seismological Society of America*, *81*(5), 1945-1957.
6. 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.
7. Ohmi, S., Watanabe, K., Shibutani, T., Hirano, N., & Nakao, S. (2002). The 2000 Western Tottori Earthquake. *Earth, planets and space*, *54*(8), 819-830.
8. Nozu, A., Nagao, T., & Yamada, M. (2007). Site amplification factors for strong-motion sites in Japan based on spectral inversion technique and their use for strong-motion evaluation. *Journal of Japan Association for Earthquake Engineering*, *7*(2), 215-234.
9. Goto, H., Sawada, S., Morikawa, H., Kiku, H., & Ozalaybey, S. (2005). Modeling of 3D subsurface structure and numerical simulation of strong ground motion in the Adapazari basin during the 1999 Kocaeli earthquake, Turkey. *Bulletin of the Seismological Society of America*, *95*(6), 2197-2215.
10. Bradley, B. A. (2014). Site-specific and spatially-distributed ground-motion intensity estimation in the 2010–2011 Canterbury earthquakes. *Soil Dynamics and Earthquake Engineering*, *61*, 83-91.
11. Jeong, S., & Bradley, B. A. (2017). Amplification of strong ground motions at Heathcote Valley during the 2010–2011 Canterbury earthquakes: Observation and 1D site response analysis. *Soil Dynamics and Earthquake Engineering*, *100*, 345-356.
12. Goto, H., & Morikawa, H. (2012). Ground motion characteristics during the 2011 off the Pacific coast of Tohoku earthquake. *Soils and Foundations*, *52*(5), 769-779.

13. Goda, K., Kiyota, T., Pokhrel, R. M., Chiaro, G., Katagiri, T., Sharma, K., & Wilkinson, S. (2015). The 2015 Gorkha Nepal earthquake: insights from earthquake damage survey. *Frontiers in Built Environment*, *1*, 8.
14. Rajaure, S., Asimaki, D., Thompson, E. M., Hough, S., Martin, S., Ampuero, J. P., Dhital, M. R., Inbal, A., Takai, N., Shigefuji, M., Bijukchhen, S., Ichiyanagi, M., Sasatani, T., & Paudel., L. (2017). Characterizing the Kathmandu Valley sediment response through strong motion recordings of the 2015 Gorkha earthquake sequence. *Tectonophysics*, *714*, 146-157.
15. Tallett-Williams, S., Gosh, B., Wilkinson, S., Fenton, C., Burton, P., Whitworth, M., Datla, S., Franco, G., Trieu, A., Dejong, M., Novellis, V., White, T., & Lloyd, T. (2016). Site amplification in the Kathmandu Valley during the 2015 M7. 6 Gorkha, Nepal earthquake. *Bulletin of Earthquake Engineering*, *14*(12), 3301-3315.
16. McGowan, S. M., Jaiswal, K. S., & Wald, D. J. (2017). Using structural damage statistics to derive macroseismic intensity within the Kathmandu valley for the 2015 M7. 8 Gorkha, Nepal earthquake. *Tectonophysics*, *714*, 158-172.
17. Yamanaka, H., Chimoto, K., Miyake, H., Tsuno, S., & Yamada, N. (2016). Observation of earthquake ground motion due to aftershocks of the 2016 Kumamoto earthquake in damaged areas. *Earth, Planets and Space*, *68*(1), 197.
18. Yamada, M., Ohmura, J., & Goto, H. (2017). Wooden building damage analysis in Mashiki Town for the 2016 Kumamoto earthquakes on April 14 and 16. *Earthquake spectra*, *33*(4), 1555-1572.
19. Campbell, K. W. (1976). A note on the distribution of earthquake damage in Long Beach, 1933. *Bulletin of the Seismological Society of America*, *66*(3), 1001-1005.
20. Trifunac, M. D. (2003). Nonlinear soil response as a natural passive isolation mechanism. Paper II. The 1933, Long Beach, California earthquake. *Soil Dynamics and Earthquake Engineering*, *23*(7), 549-562.
21. Milne, W. G., & Davenport, A. G. (1969). Distribution of earthquake risk in Canada. *Bulletin of the Seismological Society of America*, *59*(2), 729-754.
22. Cornell, C. A. (1968). Engineering seismic risk analysis. *Bulletin of the seismological society of America*, *58*(5), 1583-1606.
23. McGuire, R. K. (1976). *FORTRAN computer program for seismic risk analysis* (No. 76-67). US Geological Survey.
24. Frankel, A. (1995). Mapping seismic hazard in the central and eastern United States. *Seismological Research Letters*, *66*(4), 8-21.
25. Frankel, A.D., Mueller, C.S., Barnhard, T.P., Leyendecker, E.V., Wesson, R.L., Harmsen, S.C., Klein, F.W., Perkins, D.M., Dickman, N.C., Hanson, S.L. and Hopper, M.G., 2000. USGS national seismic hazard maps. *Earthquake spectra*, *16*(1), pp.1-19.

26. Fujiwara, H., Kawai, S., Aoi, S., Ishii, T., Okumura, T., Hayakawa, Y., Morikawa, N., Senna, S., Kobayashi, K. and Hao, K.X.S.: Japan seismic hazard information station, J-SHIS, *Proceedings of the 8th US National Conference on Earthquake Engineering,*2006.
27. Slejko, D., Peruzza, L., & Rebez, A. (1998). Seismic hazard maps of Italy.
28. Musson, R. M. (1999). Probabilistic seismic hazard maps for the North Balkan region. *Annals of Geophysics*, *42*(6).
29. Adams, John, and Gail Atkinson. "Development of seismic hazard maps for the proposed 2005 edition of the National Building Code of Canada." *Canadian Journal of Civil Engineering* 30, no. 2 (2003): 255-271.
30. Giardini, D., Grünthal, G., Shedlock, K. M., & Zhang, P. (1999). The GSHAP global seismic hazard map. *Annals of Geophysics*, *42*(6).
31. Crowley, H. R. M. N. S., Pinho, R., Pagani, M., & Keller, N. (2013). Assessing global earthquake risks: the Global Earthquake Model (GEM) initiative. In *Handbook of seismic risk analysis and management of civil infrastructure systems* (pp. 815-838). Woodhead Publishing.
32. Pagani, Marco, Julio Garcia-Pelaez, Robin Gee, Kendra Johnson, Valerio Poggi, Vitor Silva, Michele Simionato et al. "The 2018 version of the global earthquake model: hazard component." *Earthquake Spectra* 36, no. 1_suppl (2020): 226-251.
33. Silva, Vitor, Desmond Amo-Oduro, Alejandro Calderon, Catarina Costa, Jamal Dabbeek, Venetia Despotaki, Luis Martins et al. "Development of a global seismic risk model." *Earthquake Spectra* (2020): 8755293019899953.
34. Wald, D. J., Worden, B. C., Quitoriano, V., & Pankow, K. L. (2005). *ShakeMap manual: technical manual, user's guide, and software guide* (No. 12-A1).
35. Bazzurro, P., & Cornell, C. A. (2004). Nonlinear soil-site effects in probabilistic seismic-hazard analysis. *Bulletin of the Seismological Society of America*, *94*(6), 2110-2123.
36. International Society for Soil Mechanics and Geotechnical Engineering (ISSMGE) (1999). Manual for Zonation on Seismic Geotechnical Hazard (Revised Version). Technical Committee for earthquake geotechnical engineering, TC4. *The Japanese Geotechnical Society, Tokyo*.
37. Stewart, J. P., Klimis, N., Savvaidis, A., Theodoulidis, N., Zargli, E., Athanasopoulos, G., Pelekis, P., Mylonakis, G., & Margaris, B. (2014). Compilation of a local VS profile database and its application for inference of VS 30 from geologic-and terrain-based proxies. *Bulletin of the Seismological Society of America*, *104*(6), 2827-2841.
38. Foti, S., Hollender, F., Garofalo, F., Albarello, D., Asten, M., Bard, P. Y., Comina, C., Cornou, C., Cox, B., Di Giulio, G., Forbriger, T., Hayashi, K., Lunedei, E., Martin, A., Mercerat, D., Ohrnberger, M., Poggi, V., Renailer, F., Scilia, D., & Forbriger, T. (2018). Guidelines for the good practice of surface wave analysis: A product of the InterPACIFIC project. *Bulletin of Earthquake Engineering*, *16*(6), 2367-2420.

39. Allen, T. I., & Wald, D. J. (2009). On the use of high-resolution topographic data as a proxy for seismic site conditions (VS 30). *Bulletin of the Seismological Society of America*, *99*(2A), 935-943.
40. Wald, D. J., & Allen, T. I. (2007). Topographic slope as a proxy for seismic site conditions and amplification. *Bulletin of the Seismological Society of America*, *97*(5), 1379-1395.
41. Goto, H., Morikawa, H., Inatani, M., Ogura, Y., Tokue, S., Zhang, X.R., Iwasaki, M., Araki, M., Sawada, S. & Zerva, A. (2012). Very dense seismic array observations in Furukawa district, Japan, *Seismol. Res. Lett.*, 83(5), 765-774.
42. EPA, U.S. (2011). Exposure Factors Handbook 2011 Edition (Final Report). US Environmental Protection Agency, Washington, DC, EPA/600/R-09/052F.
43. Wakamatsu, K., & Matsuoka, M. (2013). Nationwide 7.5-arc-second Japan engineering geomorphologic classification map and Vs30 zoning. Journal of Disaster Research, 8(5), 904-911.
44. Matsuoka, M. and Wakamatsu, K. (2008): "Site Amplification Capability Map based on the 7.5-arc-second Japan Engineering Geomorphologic Classification Map", National Institute of Advanced Industrial Science and Technology, Intellectual property management, No.H20PRO-936.
45. Fujimoto, K., & Midorikawa, S. (2006). Relationship between average shear-wave velocity and site amplification inferred from strong motion records at nearby station pairs. Journal of Japan Association for Earthquake Engineering, 6(1), 11-22.
46. J-SHIS Map: http://www.j-shis.bosai.go.jp/map/?lang=en (accessed on 2020.10.07)
47. Harrower, M. (2003, November). Representing uncertainty: Does it help people make better decisions. In *Ucgis workshop: Geospatial visualization and knowledge discovery workshop* (pp. 18-20).
48. Brodlie, K., Osorio, R. A., & Lopes, A. (2012). A review of uncertainty in data visualization. In *Expanding the frontiers of visual analytics and visualization* (pp. 81-109). Springer, London.
49. Kuehn, N. M., & Scherbaum, F. (2016). A partially non-ergodic ground-motion prediction equation for Europe and the Middle East. *Bulletin of Earthquake Engineering*, *14*(10), 2629-2642.
50. Ordaz, M., & Arroyo, D. (2016). On uncertainties in probabilistic seismic hazard analysis. *Earthquake Spectra*, *32*(3), 1405-1418.
51. De Risi, R., De Luca, F., Gilder, C. E., Pokhrel, R. M., & Vardanega, P. J. (2020). The SAFER geodatabase for the Kathmandu valley: Bayesian kriging for data-scarce regions. Earthquake Spectra, 8755293020970977
52. Rodriguez-Marek, A., Bommer, J. J., Youngs, R. R., Crespo, M. J., Stafford, P. J., & Bahrampouri, M. (2020). Capturing epistemic uncertainty in site response. *Earthquake Spectra*, 8755293020970975.
53. Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

# CHAPTER 2

# BAYESIAN HIERARCHICAL MODELING OF SPATIAL DATA

## 2.1 Introduction

In this chapter, the mathematical background of a Bayesian hierarchical modelling of spatial data is introduced.

## 2.2 Bayesian inference

Bayesian inference is a way of making statistical estimations by assigning subjective probabilities to data distributions. These subjective probabilities are also known as the prior probabilities. Observations update the prior probabilities based on *Bayes' rule*. These updated or revised probabilities form the so-called posterior probabilities. Bayesian inference, thus, allows estimations to be made on data using probability models. The characteristic of Bayesian models is their explicit use of probability to quantify uncertainty in inferences based on statistical data analysis.

The foundation of Bayesian inference lies in *Bayes' rule* **[1]** which is a rule for computing conditional probabilities. Let $p(A)$ and $p(B)$ be the probabilities of two events $A$ and $B$. Also, let $p(A|\ B)$ denote the conditional probability of $A$ given $B$ and $p(B|\ A)$ be the conditional probability of $B$ given $A$.

*Bayes' rule* states that

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} \tag{2.1}$$

In Bayesian terminology, $p(A)$ is the prior probability, $p(B|A)$ is called the conditional probability or likelihood, $p(B)$ is the marginal probability and $p(A|B)$ is called the posterior probability.

Gelman et al. **[2]** idealizes the process of Bayesian data analysis into the following three steps:

1) The first step is the setting up of a full probability model which is a joint probability distribution of all known and unknown parameters in a problem.
2) The second step is calculating and interpreting the posterior distribution i.e. the conditional probability distribution of the unknown parameters of interest given the known parameters.
3) The third step is understanding the implications of the resulting posterior distribution.

By modelling both the observed data and any unknowns as random variables, the Bayesian approach to statistical analysis provides a coherent framework for combining complex data models and external knowledge or expert opinion **[3]**. Let $\theta$ denote unknown parameters of interest and $y=(y_1,y_2,\ldots\ldots,y_n)$ denote the observed data. To make probability statements about $\theta$ given $y$, we must begin with a model providing a joint probability distribution for $\theta$ and $y$. The joint probability mass or density function can be written as a product of two densities; the prior distribution $p(\theta)$ and the data distribution $p(y|\ \theta)$ as

$$p(\theta,y)=p(\theta)p(y\mid\theta) \tag{2.2}$$

Simply conditioning on the known value of the data $y$, using *Bayes' rule*, the posterior density becomes:

$$p(\theta\mid y)=\frac{p(\theta,y)}{p(y)}=\frac{p(\theta)p(y\mid\theta)}{p(y)} \tag{2.3}$$

where

$$p(y)=\int p(y,\theta)d\theta=\int p(\theta)p(y\mid\theta)d\theta \tag{2.4}$$

An equivalent form of **equation (2.3)** yielding the un-normalized posterior density is given by

$$p(\theta\mid y)\propto p(\theta)p(y\mid\theta) \tag{2.5}$$

**Equation (2.2) ~ (2.5)** summarizes the fundamental idea of Bayesian inference; the primary task of any specific application of Bayesian analysis is to develop a model $p(\theta,y)$ and perform the relevant computations to summarize $p(\theta|\ y)$ in appropriate ways.

**Equation (2.3)** says that Bayes' rule with some probability model implies that the data $y$ affect the posterior inference only through the function $p(y|\ \theta)$. This function when considered as a function of $\theta$ for a fixed $y$, is called the likelihood function.

**Equation (2.4)** is often termed as the marginal distribution of $y$, but better known as prior predictive distribution in this case: prior because it is not conditional on a previous observation of the process and predictive because it is the distribution for a quantity that is observable **[2]**.

## 2.3 Single and two-parameter Bayesian models

### 2.3.1 Single parameter model

In section 2.2, the statistical model had only one parameter to be estimated i.e. $\theta$ is one-dimensional. These models are called single parameter models. An example of a single parameter model can be a normal distribution with an unknown mean, $\mu$ and known variance, $\sigma^2$. The likelihood is given by **[4]**

$$p(y \mid \theta) = p(y \mid \mu, \sigma^2) = \prod_{i=1}^{n} p(y_i \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\} \tag{2.6}$$

Using

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.7}$$

and

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \tag{2.8}$$

**equation (2.6)** can be rewritten as **[22]**

$$p(y \mid \mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}[ns^2 + n(\bar{y} - \mu)^2]\right) \tag{2.9}$$

Assuming $\sigma^2$ to be a constant,

$$p(y \mid \mu) \propto \exp\left(-\frac{n}{2\sigma^2}(\bar{y} - \mu)^2\right) \propto N\left(\bar{y} \mid \mu, \frac{\sigma^2}{n}\right) \tag{2.10}$$

The natural conjugate prior which has the same form as the likelihood can be written as

$$p(\mu) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \propto N\left(\mu \mid \mu_0, \sigma_0^2\right) \tag{2.11}$$

where $N(\mu_0, \sigma_0^2)$ is a normal distribution with mean $\mu_0$ and variance $\sigma_0^2$ .

Using Bayes' rule, the posterior is given by **[4]**

$$p(\mu \mid y) \propto p(y \mid \mu,\sigma)p(\mu \mid \mu_0,\sigma_0^2) \tag{2.12}$$

$$p(\mu \mid y) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_i (y_i-\mu)^2\right] \times \exp\left[-\frac{1}{2\sigma_0{}^2}(\mu-\mu_0)^2\right] \tag{2.13}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum (y_i^2+\mu^2-2y_i\mu)+\frac{-1}{2\sigma_0^2}(\mu^2+\mu_0^2-2\mu_0\mu)\right] \tag{2.14}$$

As the product of two normal distributions is a normal distribution, equation **(2.14)** can be rewritten as

$$p(\mu \mid y) \propto \exp\left[\frac{-\mu^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)+\mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_i y_i}{\sigma^2}\right)-\left(\frac{\mu_0^2}{2\sigma_0^2}+\frac{\sum_i y_i^2}{2\sigma^2}\right)\right] \tag{2.15}$$

Let the posterior distribution be a normal distribution with $N(\mu_n,\sigma_n^2)$ with mean $\mu_n$ and variance $\sigma_n^2$. So, $p(\mu|y)$ becomes

$$p(\mu \mid y) \propto \exp\left[-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right] \tag{2.16}$$

Equating **equations (2.15)** and **(2.16)**, and matching the coefficients of $\mu^2$ **[4],** we have

$$\sigma_n^2 = \frac{1}{(\frac{n}{\sigma^2})+(\frac{1}{\sigma_0^2})} \tag{2.17}$$

Equating equations **(2.15)** and **(2.16)**, and matching the coefficients of $\mu$ **[4]**, we have

$$\mu_n = \sigma_n^2\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\bar{y}}{\sigma^2}\right) \tag{2.18}$$

Thus, Bayesian inference results in the estimation of the unknown posterior mean $\mu_n$ and variance $\sigma_n^2$ of the normal distribution. The calculation was simple as a conjugate prior was available.

### 2.3.2 Two parameter model

In this section, we consider a two-dimensional unknown vector $\theta$. An example can be a normal distribution with both an unknown mean, $\mu$ and an unknown variance, $\sigma^2$. The likelihood is given by **[5]**

$$p(y \mid \theta) = p(y \mid \mu, \sigma^2) = \prod_{i=1}^{n} p(y_i \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \tag{2.19}$$

Now, in this case as there are two unknown parameters, we apply a hierarchical prior. Hierarchical models are explained in detail in section 2.4.

The prior assigned to the mean, is conditional on the variance and is a normal distribution given by

$$p(\mu \mid \sigma^2) = (2\pi \frac{\sigma^2}{\nu})^{-1/2} \exp\left( -\frac{1}{2} \frac{\nu}{\sigma^2} (\mu - \mu_0)^2 \right) \tag{2.20}$$

where $\mu$ is assumed to be normally distributed with mean $\mu_0$ and variance $\tau_0^2 = \frac{\sigma^2}{\nu}$

The prior assigned to the variance is an Inverse-Gamma distribution given by

$$p(\sigma^2) = \frac{(k\sigma_0^2)^{k/2}}{2^{k/2}\Gamma(\frac{k}{2})} \frac{1^{\frac{k}{2}+1}}{\sigma^2} \exp(-\frac{1}{2} \frac{k\sigma_0^2}{\sigma^2}) \tag{2.21}$$

where $k$ and $1/\sigma_0^2$ are the parameters of the inverse Gamma distribution **[5]**.

The posterior distribution of the mean conditional on the variance is given by

$$p(\mu \mid y, \sigma^2) = (2\pi\tau_n^2)^{-1/2} \exp(-\frac{1}{2\tau_n^2} (\mu - \mu_n)^2) \tag{2.22}$$

where

$$\mu_n = \frac{1}{n+\nu} \left[ \sum_{i=1}^{n} y_i + \nu\mu_0 \right] \tag{2.23}$$

$$\tau_n^2 = \frac{\sigma^2}{n+\nu} \tag{2.24}$$

The analytical estimation of the posterior distribution of the mean is a little complicated and has been skipped in this section. The reason of showing the process of Bayesian inference for a two-parameter model is that as the unknown parameters increase the algebra required for the analytic derivation of posterior distributions become more and more complicated. Also, in the above examples, only normal distributions are considered for the analysis, however, as the probability distributions deviate from the usual ones, the calculations become more complex. Bayesian analysis of realistic probability models are too cumbersome for practical applications.

## 2.4 Hierarchical Bayesian model

Many statistical problems involve multiple parameters that can be regarded as mutually connected in some ways depending on the structure of the problem at hand. These problems can be naturally modelled hierarchically, with observable outcomes modeled conditionally on certain parameters, which in turn are given a probabilistic distribution in terms of further parameters known as hyper parameters **[2]**. In practice, many times nonhierarchical models are inappropriately used for data which can be better modeled using hierarchical models. Using only a few parameters, nonhierarchical models cannot fit large datasets accurately whereas the use of too many parameters tend to over fit the dataset leading to inferior predictions for the new dataset. In contrast, a hierarchical model can have enough parameters to fit the data well using a probabilistic distribution to structure some dependence between some parameters and thus, in the process, avoid the problem of overfitting.

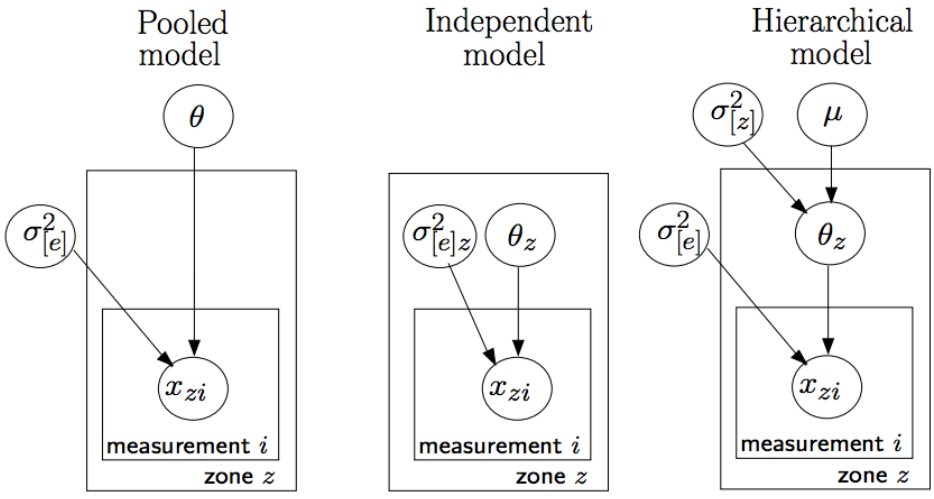



**Figure 2.1** Hierarchical and non-hierarchical approach to data modeling **[6]**

**Fig. 2.1** compares a hierarchical model with non-hierarchical models (pooled and independent models) where $\theta_z$ is the mean parameter for zone $z$, $\mu$ is the overall mean across all the zones, $\mu^2_{[z]}$ is the between zone variance and $\mu^2_{[e]}$ is the residual variance **[6]**. In a pooled model, where the parameters are treated only at group levels, the individuality of the zones is lost as only the mean value is used. In an independent model, the correlations are lost as the zones are considered separate and non-connected. However, hierarchical model brings in the best of the pooled and the independent models.

In hierarchical Bayesian modeling, in addition to specifying the distributional model $p(y|\theta)$ for the observed data $y=(y_1,y_2,\ldots\ldots,y_n)$ given a vector of unknown parameters $\theta=(\theta_1,\theta_2,\ldots\ldots,\theta_n)$, the prior distribution is now $p(\theta|\lambda)$, where $\lambda$ is a vector of hyper parameters.

If $\lambda$ is known, inference on $\theta$ is based on its posterior distribution,

$$p(\theta \mid y,\lambda) = \frac{p(y,\theta \mid \lambda)}{p(y,\lambda)} = \frac{p(y,\theta \mid \lambda)}{\int p(y,\theta \mid \lambda)d\theta} = \frac{p(y \mid \theta)p(\theta \mid \lambda)}{\int p(y \mid \theta)p(\theta \mid \lambda)d\theta} \tag{2.25}$$

However, in practice, $\lambda$ is an unknown and so a hyperprior (a second stage) distribution $p(\lambda)$ is introduced. **Equation (2.25)** will be replaced by

$$p(\theta \mid y,\lambda) = \frac{p(y,\theta)}{p(y)} == \frac{\int p(y \mid \theta)p(\theta \mid \lambda)p(\lambda)d\lambda}{\int p(y \mid \theta)p(\theta \mid \lambda)p(\lambda)d\theta d\lambda} \tag{2.26}$$

Implicit in **equation (2.26)** is a hierarchical structure with three levels of distributional specification with primary interest in the $\theta$ level.

In **Fig. 2.2**, Sheldrake **[7]** shows an example of a hierarchical Bayesian model for event probabilities $(\theta_{Ij})$. Here the data is $(x_{Ij},n_i)$ with prior parameters $(\phi,\psi)$ and hyperprior parameters $(\mu,\tau,K,J)$. The index $i$ relates to each site with a total of $V$ sites in the model and $j$ relates to each mutually exclusive event from a total of $J$ possible outcomes.

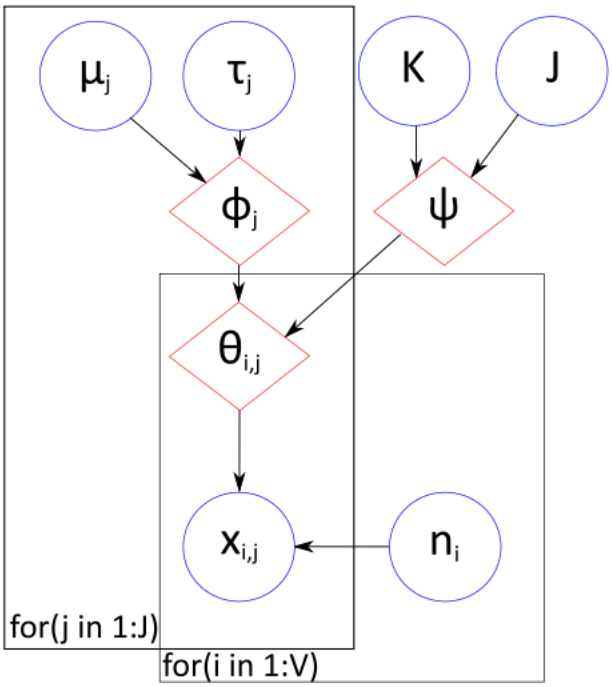



**Figure 2.2** A directed acyclic graph showing Hierarchical Bayesian model **[7]**

## 2.5 Bayesian inference using MCMC

As discussed in section 2.3, Bayesian estimation for a two-parameter model is not so simple. A computational challenge in applying hierarchical Bayesian methods is that for most realistic problems, the integrations required to do inference under **equation (2.25)** - (**2.26)** are generally not tractable in closed form and thus must be approximated numerically. Conjugate priors that enable at least partial analytic evaluation as shown in section 2.2 may often be found, but in the presence of nuisance parameters (typically unknown variances), some intractable integrations remain. However, thanks to the rapidly developing computational power over the last few decades, we have some powerful computing tools at our expense, particularly the Markov Chain Monte Carlo (MCMC) integration methods, such as the Metropolis-Hastings algorithm [**8,9**] and the Gibbs Sampler [**10,11**].

MCMC is a general sampling method based on drawing the values of $\theta$ from approximate distributions and then correcting those draws to better approximate the target posterior distribution, $p(\theta|y)$. The samples are drawn sequentially, with the distribution of the sampled draws depending on the last value drawn and thus, forming a Markov Chain. The highlight of the MCMC sampling method is not the Markov property but the point that the approximate distributions are improved at each step in the simulation, in the sense of converging to the target distribution.

## 2.6 Spatial structure: CAR model

Banerjee et al**. [3]** presents an elegant discussion on the hierarchical modelling of spatial data. However, for the point-referenced models i.e. models with spatial data at points, in all the cases, the spatial structure has been introduced using semi-variograms **[12,13]**. In this study, we are interested in modeling the distribution of the variation of the mean value around a site. A semi-variogram doesn't fit our requirements. In general, Conditional auto-regression (CAR) [**14,15,16**] which introduce spatial dependence based on a neighborhood matrix, is used only in areal data cases expect a few occasions **[17]**. In this study, we introduce the spatial structure in the hierarchical modeling using a CAR model. Its application in Bayesian analysis of hierarchical spatial models has been quite recent and mostly used for areal data models. **Fig. 2.3** illustrates a CAR model using the same Bayesian hierarchical model shown in **Fig. 2.2**.

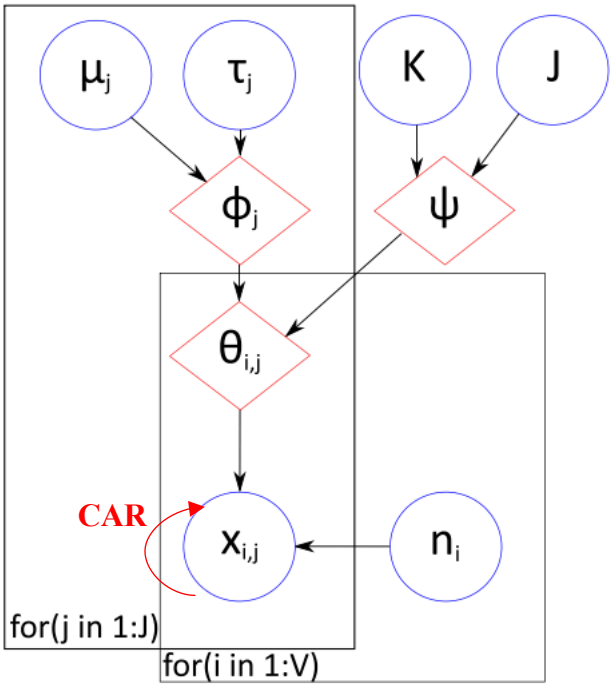



**Figure 2.3** CAR prior in a hierarchical Bayesian model (adapted from **[25]**)

The CAR model as shown in **Fig 2.3** can be written as a conditional probability given by

$$p(x_i \mid x_1, x_2, ...., x_{i-1}, x_{i+1}, x_{i+2}, ...., x_n) \tag{2.27}$$

which means that the parameter value at $x_i$ is conditional on all other values in the neighborhood of $x_i$. Conditional autoregressive models thus allow the value at a site to be conditional on the neighboring sites.

## 2.7 Model evaluation

Model evaluation is an integral part in the process of model development. It helps us to find the best model that represents the data and how well the chosen model will work in the future. Evaluating the performance of a model with all the available data for training is not acceptable as it can easily generate overoptimistic and over fitted models.

### 2.7.1 Cross-validation

Cross-validation is a model evaluation technique which partitions the original data sample into a training set to train the model and a testing set used to evaluate the model. Suppose our dataset $y$ consists of $n$ values $(y_1, y_2, \ldots\ldots\ldots, y_n)$. Instead of training a predictor to the full dataset $y$, we can train it with only the first $(n\text{-}1)$ values $(y_1, y_2, \ldots\ldots\ldots, y_{n-1})$, and then use this to form a prediction $\pi_n$ for the last value $n$.

Since I have the value of $y_n$, the quality of the prediction can be assessed by crosschecking it with some estimator variable. As $y_n$ was not a part of the training dataset, this assessment should be immune to re-substitution bias. The core idea of cross-validation **[18,19,20]** is to repeat the procedure but withholding a different value each time.

In $k$-fold cross-validation **[21,22]**, the original sample is randomly partitioned into $k$ equal size subsamples. Of the $k$ subsamples, a single subsample is retained as the validation data for model testing and the remaining $k\text{-}1$ subsamples are used as training data. The cross-validation process is then repeated $k$ times ($k$ folds) with each of the $k$ subsamples used exactly once as the testing data. The $k$ results from the $k$ folds can then be averaged to produce a single estimation.

### 2.7.2 WAIC

$k$-fold cross-validation 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, an alternative is Watanabe-Akaike information criterion (WAIC) [**23**], which is an information criterion and has the form given by **equation (2.28)** [**24**]

$$\widehat{elppd}_{WAIC} = lppd - \rho WAIC \tag{2.28}$$

where $\widehat{elppd}_{WAIC}$ is the expected log pointwise predictive density, $lppd$ is the log pointwise predictive density and $\rho WAIC$ is the effective number of parameters given by two definitions of $\rho WAIC1$ and $\rho WAIC2$ as follows:

$$\rho WAIC1 = 2\sum_{i=1}^{n}(\log\left(E_{post}p(y_i|\theta)\right) - E_{post}(\log p(y_i|\theta))) \quad (2.29)$$

$$\rho WAIC2 = \sum_{i=1}^{n} var_{post}(\log p(y_i|\theta)) \quad (2.30)$$

where $p(y|\theta)$ is a measure of predictive accuracy or also known as the likelihood, $n$ is the number of observations, $E_{post}p(y_i|\theta)$ is the likelihood for $y_i$ induced by the posterior distribution $p_{post}$ and $var_{post}$ is the posterior variance. The model with the minimum $\widehat{elppd}_{WAIC}$ is considered as the best model.

## References


1. Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. Philosophical transactions of the Royal Society of London, (53), 370-418.
2. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.
3. Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.
4. Murphy, K.P.: Conjugate Bayesian analysis of the Gaussian distribution https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf (accessed on 2018.01.30)
5. StacLect lecture notes: Bayesian estimation of the parameters of the normal distribution https://www.statlect.com/fundamentals-of-statistics/normal-distribution-Bayesian-estimation (accessed on 2018.01.25)
6. Spiegelhalter, D. (2005). Introduction to Bayesian Analysis using WinBUGS, Coursework, MRC Biostatistics unit, Cambridge.
7. Sheldrake, T. (2014). Long-term forecasting of eruption hazards: A hierarchical approach to merge analogous eruptive histories. *Journal of volcanology and geothermal research*, *286*, 15-23.
8. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087-1092.
9. Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, 57(1), 97-109.
10. Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE trans. on Pattern analysis and Machine intelligence*, 6, 721-741.
11. Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, *85*(410), 398-409.
12. Matheron, G. (1963). Principles of geostatistics. *Economic geology*, *58*(8), 1246-1266.

13. Cressie, N.A.C. (1993) Statistics for spatial data, revised edition, Wiley, New York.
14. Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B (Methodological), 36(2), 192-225.
15. Wall, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of statistical planning and inference*, *121*(2), 311-324.
16. De Oliveira, V. (2012). Bayesian analysis of conditional autoregressive models. Annals of the Institute of Statistical Mathematics, 64(1), 107-133.
17. 久保拓弥 (2012) データ解析のための統計モデリング入門, 岩波書店.
18. Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, *4*, 40-79.
19. Geisser, S. (1971). The inferential use of predictive distributions. *Foundations of Statistical Inference*, 456-469.
20. Stone, M. (1974). Cross-validation and multinomial prediction. *Biometrika*, *61*(3), 509-515.
21. Geisser, S. (1975). The predictive sample reuse method with applications. Journal of the American statistical Association, 70(350), 320-328.
22. Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. *Encyclopedia of database systems*, *5*, 532-538.
23. Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, 11, 3571-3594.
24. Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, *24*(6), 997-1016.

# CHAPTER 3

# UNCERTAINTY PROJECTED MAPPING

## 3.1 Introduction

To formulate UPM, I needed a framework that can incorporate the data uncertainty at the sites. I adopted a hierarchical Bayesian model [**1**] of the spatial distribution of site responses. The framework of hierarchical Bayesian model incorporates uncertainties and allows modeling the variabilities as parameters [**2**]. In this study, a spatial distribution of the available observations is statistically modeled and then the model parameters, which includes the information of the data uncertainty at a site, are estimated. The parameter distribution is referred to as the UPM of the data.

Until at least 20 years ago, hierarchical Bayesian models were limited within a small group of theoretical probabilists and statisticians [**1**]. The computations associated with the hierarchical models were far too complex for existing computing technology. However, in the 90's, Markov Chain Monte Carlo (MCMC) methods arrived in the scene and coupled with a revolution in faster computation processors started a new era of application of Bayesian methods. Hierarchical models have since been used extensively in spatial modelling in ecological studies, social sciences, atmospheric processes, etc. [**1,2,3,4,5**]. For example, Hooten et al. [**6**] used hierarchical Bayesian model of ground flora on large domains. Corander et al. [**7**] applied Bayesian approach to spatial modelling of genetic modelling structure. Thogmartin et al. [**8**] applied hierarchical model to study of spatial distribution of bird populations. Ver Hoef & Frost [**9**] used hierarchical Bayesian model for monitoring harbor seal changes. Hierarchical Bayesian models have also been introduced in several researches in the field of earthquake engineering [**10,11,12**].

In this chapter, the theory of UPM is introduced at first. The UPM methodology is then validated with certain numerical experiments. The validated UPM is then applied to make a site response map for a case-study area in Furukawa, Japan. The results are then discussed in comparison to a conventional mapping method.

## 3.2 The proposed UPM methodology

In this section, the proposed UPM methodology is introduced. The backbone of this methodology is a hierarchical Bayesian model of site response observations with spatially correlated random effects added to the model using a conditionally autoregressive (CAR) prior [**13**]. The novelty of the proposed methodology lies in the mapping the site response observations considering the data uncertainties at the site.

### 3.2.1 Hierarchical Bayesian model of site response observations

Let $Y_{ij}$ be a random observation at site *j* during an event *i*. I define the observation $Y_{ij}$ as a site response that is the ratio of ground motion index (i.e., PGA, PGV, etc.) at site *j* to the average of the index calculated over all the available sites during an event *i*.

The conventional mapping of $Y_{ij}$ is a spatial distribution of the average value of $Y_{ij}$ at site *j*, which is given by $\sum_{i=1}^{n} \frac{Y_{ij}}{n}$, where *n* is the number of observations. However, there exists a sample variance at each site *j*, which is not at all used in conventional mapping procedures. In this study, I define the sample variance as uncertainty of the site response and incorporate that uncertainty in the mapping. I adopt a hierarchical Bayesian model, which can estimate parameters on posterior distribution in estimating the uncertainty [**1,2**].

The hierarchical Bayesian model for $Y_{ij}$ is designed as follows:

$$Y_{ij} \sim N\left(\mu_j, \sigma_j^2\right) \tag{3.1}$$

$$\mu_j = \bar{\mu} + \Delta\mu_j \tag{3.2}$$

$$\Delta\mu_j \sim N\left(\frac{\Sigma_{m \epsilon M_j} \Delta\mu_m}{N_j}, \frac{{s_j}^2}{N_j}\right) \tag{3.3}$$

where $N(\mu, \sigma^2)$ is a normal distribution with mean $\mu$ and standard deviation $\sigma$. **Equation (3.1)** states that the observation $Y_{ij}$ at site *j* during an event *i* is a normally distributed variable with unknown site-specific mean, $\mu_j$ and unknown site-specific standard deviation, $\sigma_j$.

The site-specific mean, $\mu_j$ varies from site to site. The average of $\mu_j$ across all the available sites is denoted by $\bar{\mu}$. **Equation (3.2)** states that the site-specific mean at site *j* differs from $\bar{\mu}$ by $\Delta\mu_j$, which can be termed as the spatial fluctuation of the site-specific mean.

**Equation (3.3)** models the spatial fluctuation $\Delta\mu_j$ as a normally distributed variable. $\Delta\mu_j$ is constrained to be conditional on the sites inside a neighborhood of *j*. Based on the CAR model definition by De Oliveira [**13**], we define a neighborhood around site *j* as $M_j$ and the member number of $M_j$ as $N_j$. Note

that $M_j$ does not contain $j$ itself [**13**]. The mean of the model distribution is the average of the spatial fluctuation measured from $j$ across the neighborhood $M_j$, and the standard deviation is $s_j$. In this study, $s_j$ is another uncertainty which we term as the spatial variation. **Equation (3.3)** introduces the spatial correlation in the model.

CAR models are often used to describe spatial variation of areal observations and analyze data in diverse areas such as demography, economy, epidemiology and geography [**1,14,15**].

### 3.2.2 The novel constraint $c = s\sigma$

The goal of this study is to estimate an unknown site-specific mean value which is affected by the uncertainty at a site (site-specific standard deviation). In this hierarchical model, so far, the $s_j$ parameter of the CAR model in **equation (3.3)** controls the estimation of the site-specific mean, $\mu_j$. From **equation (3.3)**, when the spatial fluctuation within the neighborhood is relatively large, $s_j$ is large and hence, the observations at the sites are spatially less correlated and $\mu_j$ values are relatively rough. However, when the spatial fluctuation within the neighborhood is relatively small, $s_j$ is low which introduces high spatial correlation and results in relatively smooth $\mu_j$ values. The $s_j$ parameter thus controls the smoothness in the spatial distribution of $\mu_j$.

In UPM, I want the mapping to be rough, i.e. to follow the observation mean, at low values of site-specific standard deviation, and I expect the mapping be smooth when the site-specific standard deviations are high. So, I introduce an additional constraint $c = s\sigma$ in the hierarchical model to relate $s_j$ and $\sigma_j$ so that $\mu_j$ will now be affected by both $s_j$ and $\sigma_j$.

$$c = s\sigma \tag{3.4}$$

This is the novelty of the proposed methodology. Now, the estimated result of $\mu_j$, which is UPM of $Y_{ij}$, will now incorporate the information of the data uncertainty. Based on this novel constraint, we can now discuss **Fig. 1.4**. At low values of site-specific standard deviation ($\sigma_j$) in **Fig. 1.4(a)**, $s_j$ becomes high resulting in less spatial correlation and hence a rough mapping of $\mu_j$ similar to conventional mapping. However, at high values of site-specific standard deviation ($\sigma_j$) in **Fig. 1.4(b)**, $s_j$ becomes low resulting in more spatial correlation and hence a smooth mapping of $\mu_j$.

**Equations (3.1) ~ (3.4)** comprises the proposed methodology for uncertainty projected mapping.

### 3.2.3 Estimating the unknown parameters $\mu$, $\sigma$ and $s$

In Bayesian analysis, the unknown parameters $\mu$, $\sigma$ and $s$ are assigned a prior distribution and the values can be estimated based on a posterior probability distribution which is calculated using the Bayes' theorem. Instead of the theoretical calculation of the posterior probability distribution, MCMC [**16**] algorithms are used to estimate the parameters. In this paper, we use statistical software WinBUGS [**17**] for the execution.

In my proposed methodology, I provide a uniform distribution as the prior for $\sigma$ (site-specific variation) and $s$ (spatial variation). For the site-specific mean ($\mu$), I provide a zero-mean normal distribution with a large variance as a prior. All the priors are non-informative so that the estimated posteriors are not sensitive to the prior information.

### 3.2.4 Estimating the model evaluation parameter *c*

As many models with different *c* can be assumed, many different mappings are possible. Hence, model evaluation needs to be done to select the best *c*-value model. I term *c* as a model evaluation parameter. Bayesian models can be evaluated and compared in several ways [**18,19,20**]. However, in this study, I preferred performing model evaluation based on the predictive accuracy [**21,22**] as I wanted the best *c*-value model to reflect the matching with the data. Cross-validation and information criteria are two approaches of estimating prediction accuracy from a fitted Bayesian model using the log-likelihood evaluated at the posterior simulations of the parameter values [**23,24,25**].

In this study, I used *k*-fold cross-validation technique for the model evaluation. The core idea of cross-validation [**26,27,28**] is to split data, once or several times; part of data (the training set) is used to train the model, and the remaining part (the testing set) is used to validate the model. In *k*-fold cross-validation [**29,30**] the dataset is partitioned into *k* subsets of equal size. The model is built *k* times, each time with one of the *k* subsets as testing set and the remaining as the training set.

Each time a model is built with a certain *c*-value, the likelihood function of the *c*-value model is estimated. The average of all the likelihoods corresponding to the different training sets in *k*-fold cross-validation for each *c*-value model is calculated and used as the model likelihood, *L(c)*. The model with the maximum averaged likelihood is preferred as the best model. However, the exact selection of the model for the a UPM plot varies case by case and is referred to a plot of variation of the model likelihoods, *L(c)* with the *c*-values. The $\mu_j$ from the selected best model are used for UPM of $Y_{ij}$ .

The likelihood at site *j* is given by

$$L_j = L\left(\mu_j, \sigma_j^2; x_1, x_2, x_3, \dots\dots x_n\right) = \left(2\pi\sigma_j^2\right)^{\frac{-n}{2}} \exp\left(\frac{-1}{2\sigma_j^2}\sum_{a=1}^{n}\left(x_a - \mu_j\right)^2\right) \quad (3.5)$$

where $\mu_j$ and $\sigma_j$ are the site-specific mean and site-specific standard deviations obtained from certain c-value model. $x_1, x_2, x_3, \ldots\ldots x_n$ are observations in the testing set of the $k$- fold cross-validation process, and $n$ is the number of observations.

The likelihood for the c-value model is given by

$$L(c) = \prod_{j=1}^{m} L_j \tag{3.6}$$

where $\mu_j$ at $j = 1,2, \ldots, m$ are treated as independent and identically distributed (*i.i.d.*).

### 3.2.5 The missing sites

In practice, measurement sites are limited, and I need to estimate the values at the missing sites to produce a map for an area. Missing sites refer to those sites where there are no recorded observations. In UPM, the values at the missing sites are naturally estimated based on the spatial structure introduced by the CAR model. This means the UPM works on as a spatial interpolation.

When available sites are not uniformly distributed and a distribution map needs to be prepared in an area, I create a uniformly distributed grid of missing sites. In this case, while defining the neighborhood, missing site locations are considered.

## 3.3 Numerical experiments

In this section, some numerical experiments designed to validate the proposed UPM methodology are discussed. The numerical experiments have been categorized into two groups; (1) In numerical experiment A, all the sites are measuring sites, and (2) in numerical experiment B, some missing sites have been introduced.

### 3.3.1 Numerical experiment A

**Fig. 3.1** shows the dataset for numerical experiment A. In this experiment, five random samples are artificially generated as the observations at 50 sites in one-dimension. The random samples follow a normal distribution. The given mean values of the normal distribution vary in sinusoidal manner along the site locations. The given standard deviations(uncertainties) gradually increase from the left to right. The standard deviation is the uncertainty we are trying to address in this study.

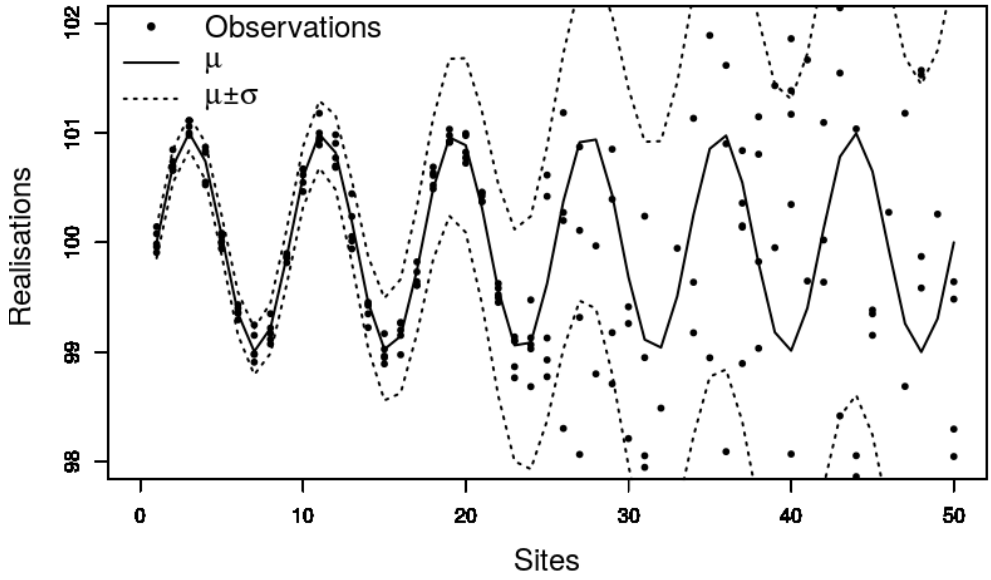



**Figure 3.1** Dataset for numerical experiment A

**Fig. 3.2** shows the conventional mapping for this dataset. The conventional mapping connects the observation means at the sites and is not sensitive to the change in uncertainty from the left to right. It should be noted that the observation mean values vary from the given mean values of the normal distribution. However, it does follow the sinusoidal curve of the given mean values along the sites.

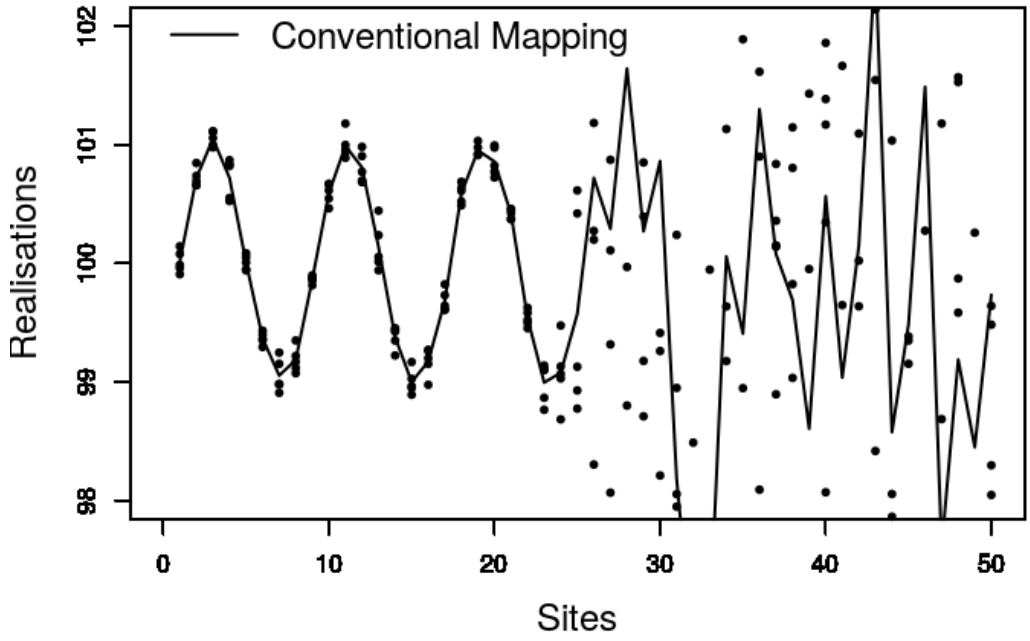



**Figure 3.2** Conventional mapping for numerical experiment A

The proposed UPM methodology as described in section 3.2 was then applied to the dataset of numerical experiment A. The spatial models differed based on the chosen *c*-values for the model. The best *c*-value model is selected based on model evaluation.

Model evaluation is done using 5-fold cross-validation method. In 5-fold cross-validation method, of the total observation data, 80% is used as training set and 20% is used as testing set for the model evaluation. **Fig. 3.3** shows one of the training results for three models with different *c*-values during model evaluation. It is observed that as the *c*-values decrease, smoothness in the mapping increases. **Fig. 3.4** shows the variation of model likelihood (in logarithm) with *c*-values. The *c*-values used in **Fig. 3.4** are highlighted using the same colors in **Fig. 3.3**. In **Fig. 3.4**, the multiple plots for a single *c*-value represent the likelihood values for different training sets. We select $c = 10^{-4}$ for the best model and the selection is indicated by an arrow in **Fig. 3.4**.

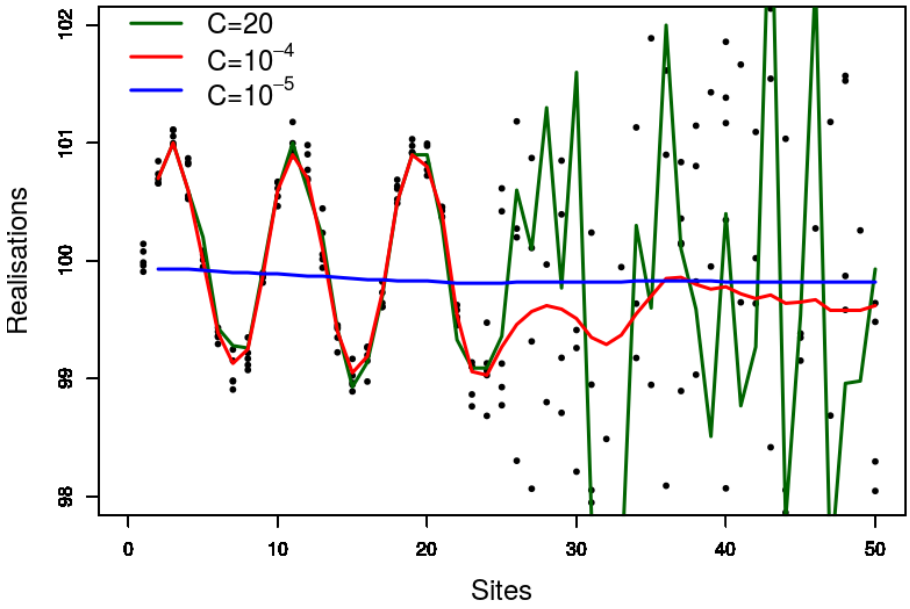



**Figure 3.3** Mapping results for a training set during the model evaluation

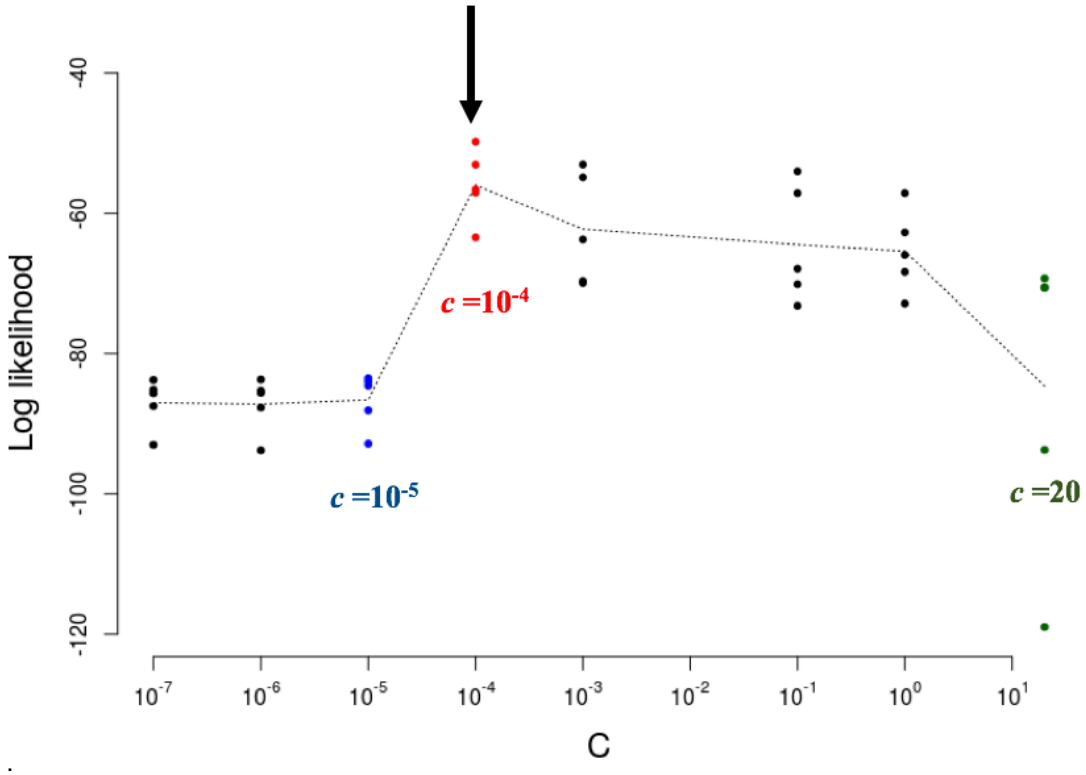



**Figure 3.4** Variation of model likelihood (in logarithm) with *c*- values. The multiple plots for a single *c*-value represent the likelihood values for different training sets. The dotted black line is the line of averaged likelihood value. The arrow indicates the selection of the best model.

I term the mapping of the site-specific mean ($\mu_j$) using the best model as the UPM of $Y_{ij}$. The red line in **Fig. 3.5** shows the UPM for the numerical experiment A using the $c = 10^{-4}$ model applied to the whole dataset. In **Fig. 3.5**, the UPM is observed to follow the conventional mapping when the site-specific standard deviation is low (on the left) and becomes smooth as the site-specific standard deviation increases towards the right. This effect is introduced due to the constraint imposed on the model by $c = s\sigma$. When $\sigma$ is low and $s$ is high, UPM is relatively rough and follows the observation mean value at the sites as the sites are spatially less correlated. As $\sigma$ increases, the $s$ decreases, leading to a high spatial correlation in the model, which introduces a smoothing effect in the mapping (**Fig. 3.5**). This effect enhances the detailed visual (mapping) on significantly different observations expected between the sites with low standard deviation and rough visual (mapping) on insignificant observations between the sites.

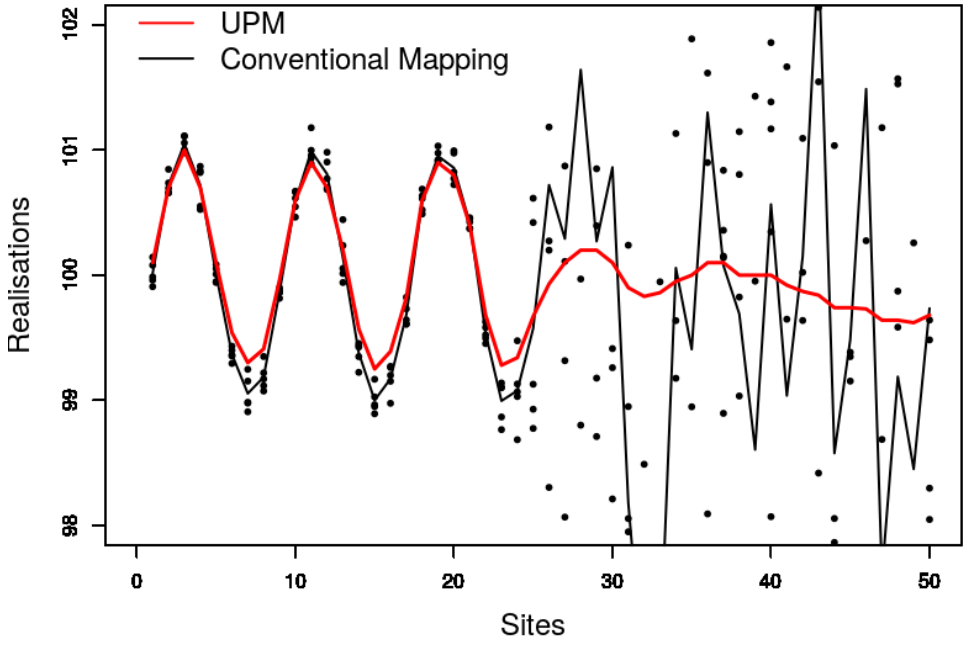



**Figure 3.5** UPM for numerical experiment A

### 3.3.2 Numerical experiment B

Numerical experiment B is designed to validate the ability of our proposed UPM methodology in estimating the values at missing sites. Two datasets are created by removing certain sites from the dataset in numerical experiment A. In the dense dataset, observations from 44% of the original sites are randomly removed, whereas in the sparse dataset, observations from 72% of the original sites are randomly removed. The dense and sparse datasets for numerical experiment B are shown in **Fig. 3.6**. As discussed in section 3.2, in UPM, the values at the missing sites are naturally estimated based on the spatial structure introduced by the CAR model.

**Figs 3.6a** and **b** show the UPM results for the numerical experiment B. In both dense and sparse datasets, UPM has captured the effect of the varying site-specific standard deviation. As in the numerical experiment A, UPM results are relatively rough and follow the observation mean values in the left where the standard deviation is low but smoothens out as the standard deviation increases in the right. Also, in **Figs. 3.6a** and **b**, UPM results are compared with a conventional mapping method called Kriging [**31**]. Kriging is a spatial interpolation algorithm and primarily used to estimate the missing site values by spatial correlations [**31,32**]. However, kriging strictly follows the observation means at the sites irrespective of the variation of the standard deviation (uncertainty). This is clearly seen in **Figs 3.6a** and **b**, where the Kriging results keep the observation means at the sites, unchanged. However, this is not the case for the UPM results. At low values of standard deviation, like Kriging results, UPM also follows the sample means. However, as the standard deviation increases in the right side, UPM results smoothen out unlike the Kriging results which still follow the observation means.

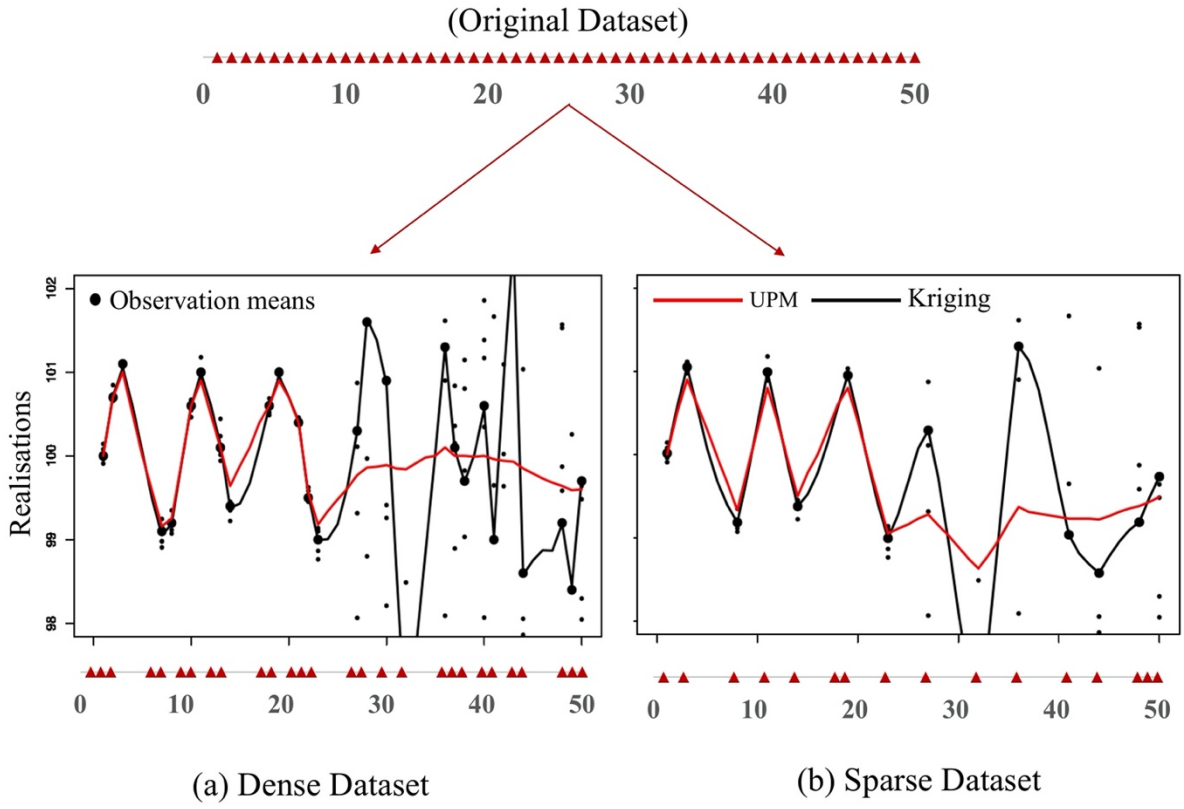



**Figure 3.6** Dataset and results for numerical experiment B

**Fig. 3.7** discusses the statistical significance of the obtained UPM and Kriging results. It uses the results shown in **Fig. 3.6(a)**. Let us draw attention to two pairs of sites. One of the pairs, A1 and A2 is from the low uncertainty part whereas the other B1 and B2 is from the high uncertainty part. Between A1 and A2, the Kriging and the UPM coincides. Whereas between B1 and B2, the Kriging and UPM doesn't agree with each other. In order to explain this situation, let us focus on the histogram of the observation data at the two pair of sites. **Fig. 3.7(a)** shows that the histograms are quite distinct in the case of pair A1 and A2. However, for the pair B1 and B2, the histograms are overlapping (**Fig. 3.7b**). Thus, UPM clearly captures the statistical significance of the mean value difference in the mapping process.

In the next section, the proposed UPM methodology is applied to real observations from a dense seismic array in a case-study area in Furukawa, Japan.

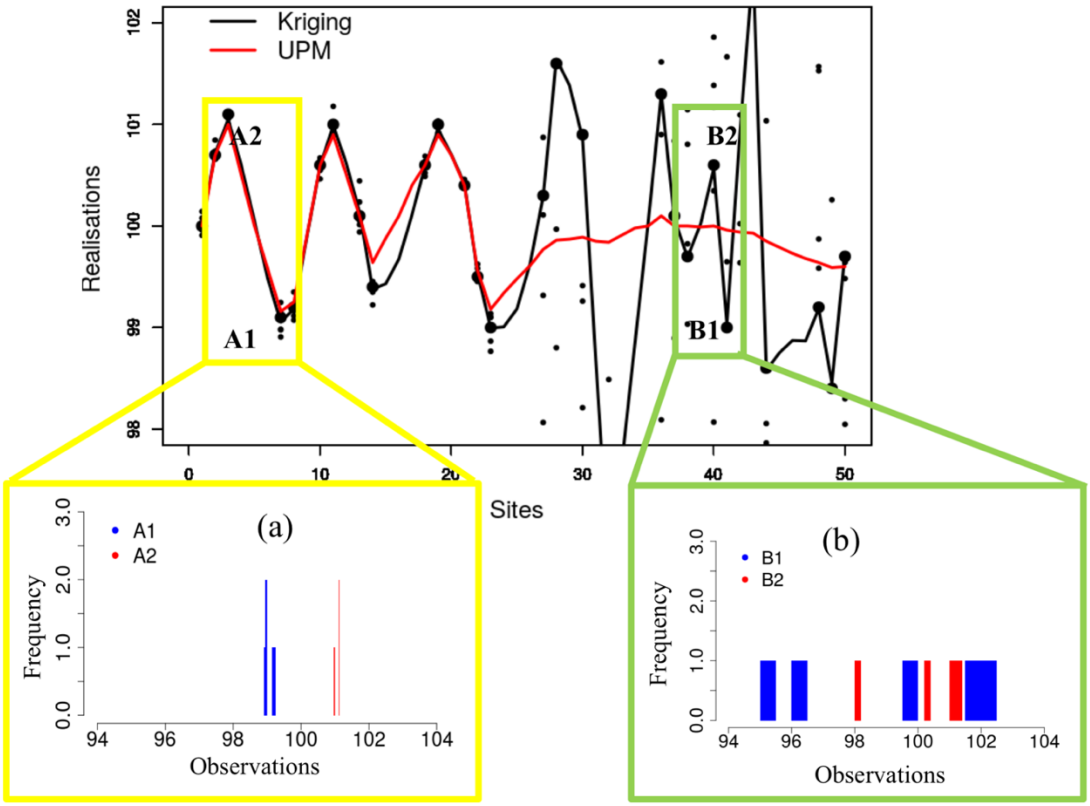
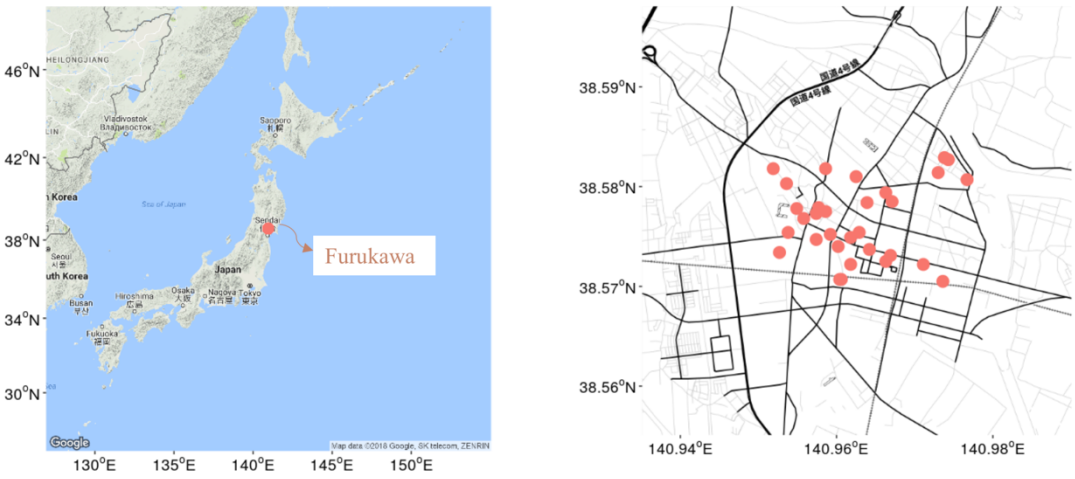



**Figure 3.7** Statistical significance of the results for numerical experiment B

## 3.4 Application: A case study in Furukawa, Japan

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 recorded severe damage in downtown residential areas. Significant spatial differences caused mainly due to site amplification were observed even in sub-kilometer scales [**33,34**]. On the aftermath of the earthquake, a very dense seismic network for strong ground motions is being operated at Osaki city. **Fig. 3.8** shows the layout of the seismic array in the significantly damaged area in Osaki city. The seismic array observation is jointly organized by Kyoto University, Tokyo Institute of Technology, Osaki city office and aLab Co. Ltd.

In this case study, earthquake data from 31 sites in the seismic array is used to generate a site response map for the area. Based on the availability of observation data at the sites, 100 earthquake events between 29$^{th}$ October,2011 and 23$^{rd}$ August,2015, were used for the analysis.




(a) (b)

**Figure 3.8** The layout of the seismic array at Furukawa district, Osaki city, Japan.

Let $Y_{ij}$ be the site response observation at site $j$ during an earthquake event $i$. In this case study, the site response $Y_{ij}$ is defined as the logarithmic ratio of ground motion index (i.e., $PGA$) at site $j$ to the average of the index calculated over all the available sites during an event $i$.

To generate a UPM map of the site responses, the dataset comprised of 431 sites with 31 measurement sites from the seismic network and 400 missing sites, all distributed in a rectangular grid.

The proposed methodology of UPM was then applied to the case study area. The best $c$-value model selected based on model evaluation i.e. 5-fold cross-validation method as discussed in section 3.2.4. **Fig. 3.9** shows the variation of model likelihood (in logarithm) with c-values. Based on the maximum likelihood and the maps produced by different $c$-value models (**Fig. 3.10**), I select $c$ =1/75 as the best model.

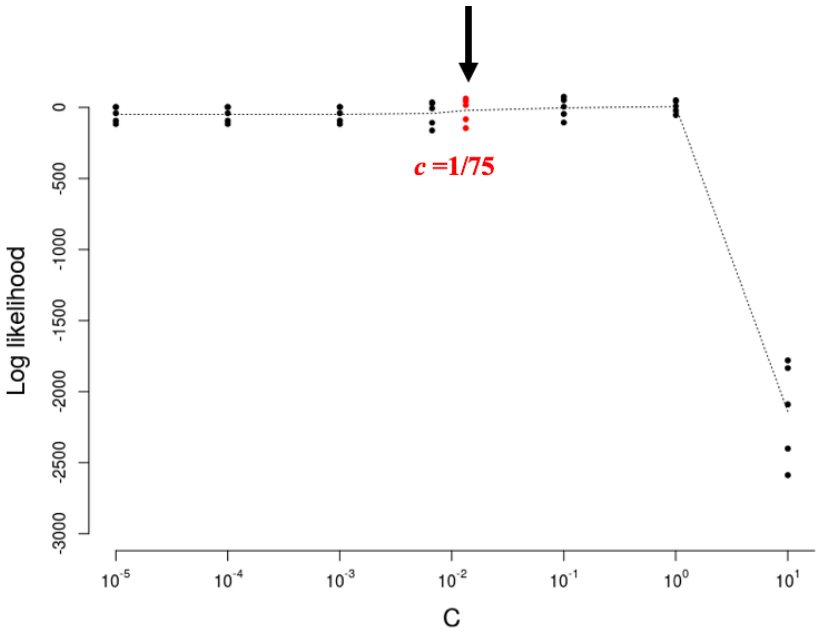



**Figure 3.9** Variation of model likelihood (in logarithm) with *c*-values for the Furukawa case study. The multiple plots for a single *c*-value represent the likelihood values for different training sets. The dotted black line is the line of averaged likelihood value. The arrow show the selection of the best model.

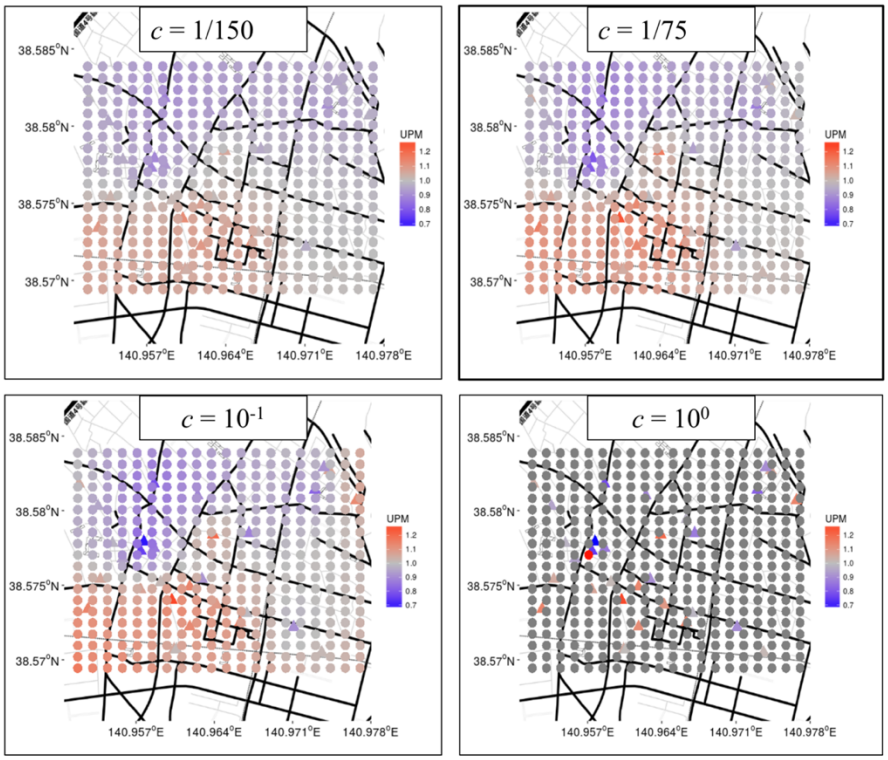



**Figure 3.10** The site response maps for some *c*-value near the best model *c* =1/75. The gray colors indicate values outside the range mentioned in the figure.

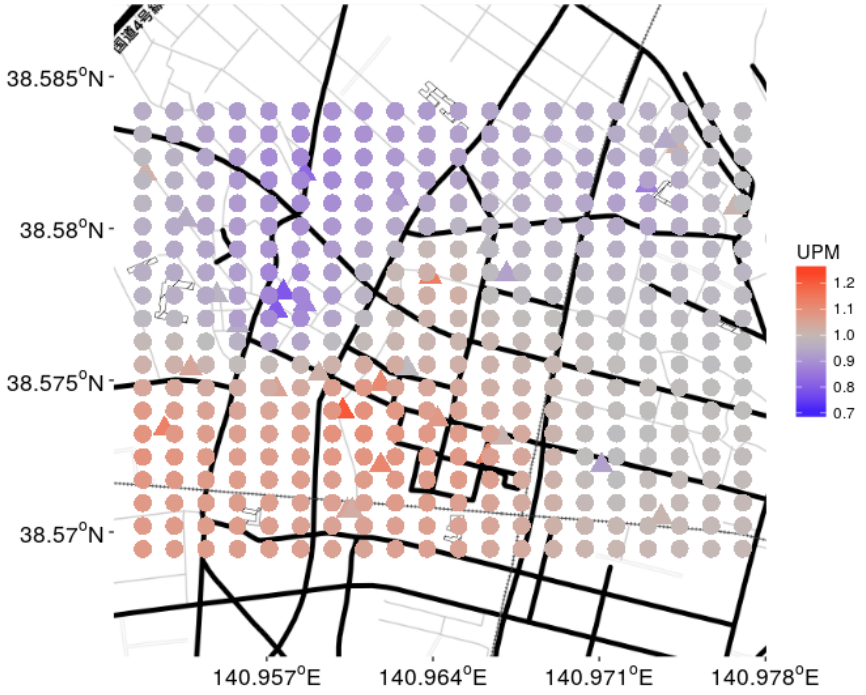



**Figure 3.11** The UPM results for the site response mapping in Furukawa district, Japan

**Fig. 3.11** shows the UPM site response map for Osaki city in Furukawa. The transition of the site response values between neighboring sites are smooth. There is no sudden occurrence of zones of extreme values. **Fig. 3.12** compares the UPM (**Fig. 3.12a**) with Kriging results (**Fig. 3.12b**) for the site response mapping. Let us focus on two pair of sites; Yellow Zone (YZ), C1 and C2 and Green Zone (GZ), D1 and D2, in both the maps. In the Kriging map, both the pairs of sites are observed to have distinct mean values; namely, YZ, C2 (in blue) and C1 (in red) and GZ, D1 (in blue) and D2 (in red). However, in the UPM map, although pair YZ, C1 and C2 can be easily distinguished, pair GZ, D1 and D2 does not appear to be very distinct. To better understand this situation, I looked at the distribution of the observation data for both the pair of sites. Inset **Fig. 3.12(c)** and **(d)** show the associated histogram of the site response observations. The histogram shows the distributions of observation data is quite distinct for pair YZ, but the data distributions cannot be distinguished clearly for pair GZ. This situation is clearly captured by UPM as at higher uncertainties the results are smoothened emphasizing the difficulty in distinguishing the sites whereas at lower uncertainties the results are rough enough to distinguish the sites clearly. UPM results are statistically significant whereas such an explanation is not possible in the case of a map produced by Kriging. In fact, this conclusion can be extended to any other mapping technique as data uncertainty is not considered in the conventional mapping processes.

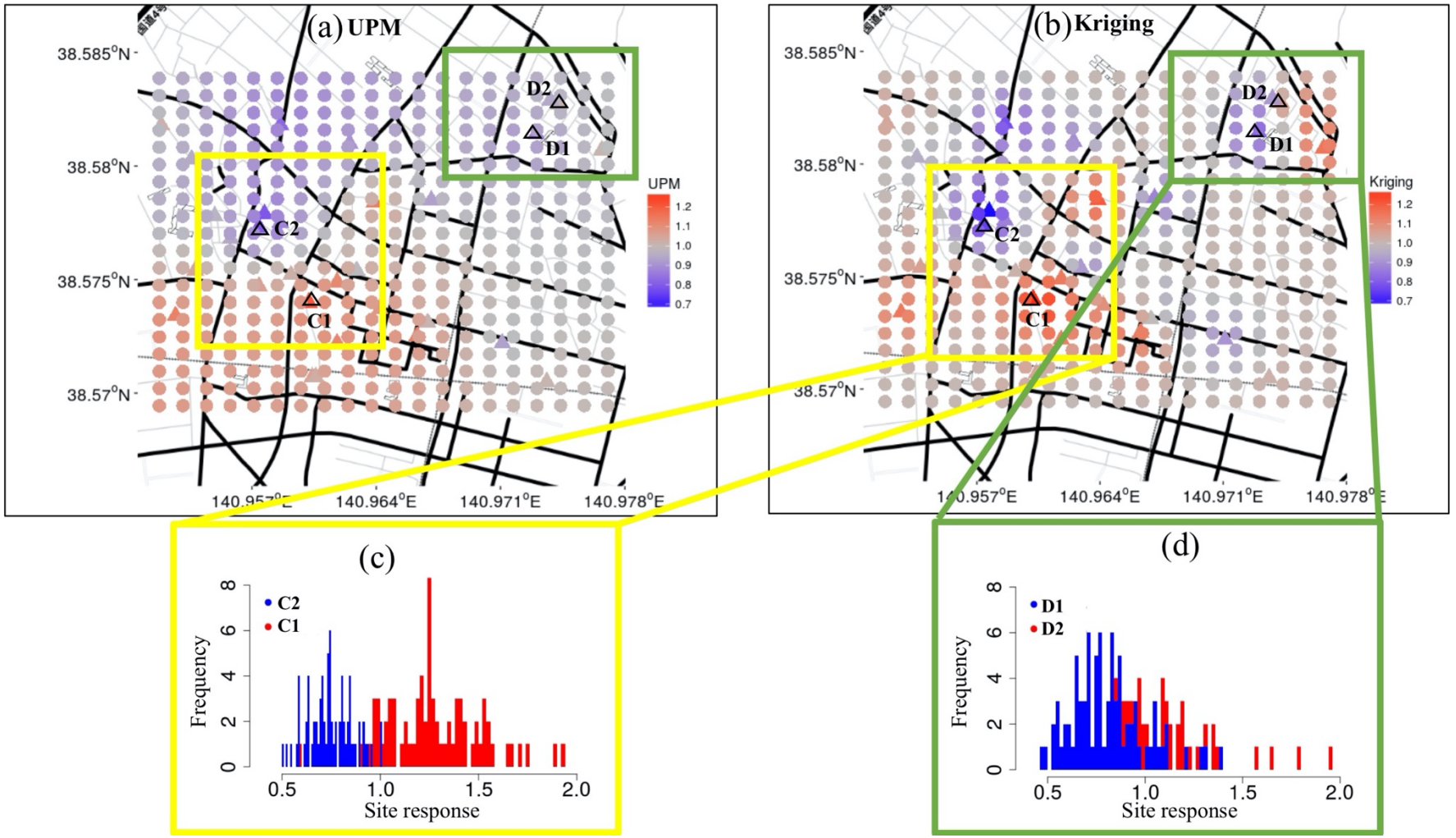



**Figure 3.12** UPM results compared with Kriging values for site response variation in Furukawa

The UPM results are observed to be smooth as compared to the conventional mapping (**Fig. 3.12**). Introducing this uncertainty-controlled smoothness in mapping was a primary objective of this research. In many conventional maps, one sometimes come across zones of extremes. The extreme values cannot be always discussed statistically and that raises a question of whether that extreme value is acceptable or not. So, I wanted to include uncertainness of the data and produce a map where the extreme values are smoothed based on the uncertainty and could be discussed statistically.

## 3.5 Conclusion

In this chapter, the problem of incorporating uncertainties in the site response mapping, has been addressed. The uncertainty primarily addressed here is the site-specific uncertainty, $\sigma$. The objective of the study had been to make the site-specific uncertainty reflect on the map resolutions of a site response mapping. To meet this objective, a methodology was proposed where the mapping parameter, mean is affected by the site-specific uncertainty. The proposed methodology of UPM which is based on a hierarchical Bayesian modeling imposed a novel constraint $c = s\sigma$ on the mapping. The result was a mapping which is rough, i.e. follows the observation mean, at low values of site-specific uncertainty, and becomes smooth at higher values of site-specific uncertainties.

To validate the proposed UPM methodology, some numerical experiments were designed in one-dimension. The mapping results from the proposed UPM methodology were found to reflect the site-specific uncertainties in the map resolutions. The detailed visual (mapping) on significantly different observations expected between the sites with low standard deviation and rough visual (mapping) on insignificant observations between the sites, was enhanced in the maps produced by the proposed UPM methodology. UPM could also generate satisfactory mapping when not all the sites had recorded observation data. The proposed methodology could spatially interpolate and estimate values at the missing sites. The values at the missing sites are naturally estimated based on the spatial structure introduced by the CAR model. Also, the mapping results were compared with a conventional mapping technique called Kriging. It was observed that unlike the UPM values which are sensitive to the variation of data uncertainty, the Kriging values are not affected by the change in data uncertainties.

Once UPM was validated with the numerical experiments, it was then applied to site-response data from a case study area in Furukawa district of Japan. On the aftermath of the 2011 off the Pacific coast of Tohoku Earthquake in Japan, a very dense seismic network for strong ground motions is being operated in Furukawa. Based on the availability of data at the sites, data from 100 earthquake events were used to prepare a site-response map of the area. The UPM results are observed to be smooth as compared to the conventional mapping and the smoothness can be discussed based on the statistical significance of the mean site response variations between the sites. Introducing this uncertainty-controlled smoothness in mapping was the primary objective of this study.

## References


1. Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.
2. Wikle, C. K. (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, *84*(6), 1382-1394.
3. Jackman, S. (2009). *Bayesian analysis for the social sciences* (Vol. 846), Chapter 7: Hierarchical Statistical Models, John Wiley & Sons.
4. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis. CRC press.
5. Wikle, C. K., Berliner, L. M., & Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, *5*(2), 117-154.
6. Hooten, M. B., Larsen, D. R., & Wikle, C. K. (2003). Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. *Landscape ecology*, *18*(5), 487-502.
7. Corander, J., Sirén, J., & Arjas, E. (2008). Bayesian spatial modeling of genetic population structure. *Computational Statistics*, *23*(1), 111-129.
8. Thogmartin, W. E., Sauer, J. R., & Knutson, M. G. (2004). A hierarchical spatial model of avian abundance with application to Cerulean Warblers. *Ecological Applications*, *14*(6), 1766-1779.
9. Ver Hoef, J. M., & Frost, K. J. (2003). A Bayesian hierarchical model for monitoring harbor seal changes in Prince William Sound, Alaska. *Environmental and Ecological Statistics*, *10*(2), 201-219.
10. Kuehn, N. M., & Scherbaum, F. (2016). A partially non-ergodic ground-motion prediction equation for Europe and the Middle East. *Bulletin of Earthquake Engineering*, *14*(10), 2629-2642.
11. Natvig, B., & Tvete, I. F. (2007). Bayesian hierarchical space–time modeling of earthquake data. *Methodology and Computing in Applied Probability*, *9*(1), 89-114.
12. Selva, J., & Sandri, L. (2013). Probabilistic seismic hazard assessment: Combining Cornell-like approaches and data at sites through Bayesian inference. *Bulletin of the Seismological Society of America*, *103*(3), 1709-1722.
13. De Oliveira, V. (2012). Bayesian analysis of conditional autoregressive models. *Annals of the Institute of Statistical Mathematics*, *64*(1), 107-133.
14. Cressie, N.A.C. (1993) Statistics for spatial data, revised edition, Wiley, New York.
15. 久保拓弥 (2012) データ解析のための統計モデリング入門, 岩波書店.
16. Gilks, W.R., Richardson, S. & Spiegelhalter, D. (1995). Markov chain Monte Carlo in practice, CRC press.
17. Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS user manual.
18. Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 1151-1172.

19. Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733-760.
20. Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, *24*(6), 997-1016.
21. Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*(365), 153-160.
22. Vehtari, A., & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. Neural computation, 14(10), 2439-2468.
23. Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and computing, 27(5), 1413-1432.
24. Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. Statistics and Computing, 27(3), 711-735.
25. Millar, R. B. (2018). Conditional vs marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation. *Statistics and Computing*, *28*(2), 375-385.
26. Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, *4*, 40-79.
27. Geisser, S. (1971). The inferential use of predictive distributions. *Foundations of Statistical Inference*, 456-469.
28. Stone, M. (1974). Cross-validation and multinomial prediction. *Biometrika*, *61*(3), 509-515.
29. Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American statistical Association*, *70*(350), 320-328.
30. Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. *Encyclopedia of database systems*, *5*, 532-538.
31. Matheron, G. (1963). Principles of geostatistics. *Economic geology*, *58*(8), 1246-1266.
32. Cressie, N. (1988). Spatial prediction and ordinary kriging. Mathematical geology, 20(4), 405-421.
33. Goto, H., & Morikawa, H. (2012). Ground motion characteristics during the 2011 off the Pacific coast of Tohoku earthquake. *Soils and Foundations*, *52*(5), 769-779.
34. Goto, H., Morikawa, H., Inatani, M., Ogura, Y., Tokue, S., Zhang, X.R., Iwasaki, M., Araki, M., Sawada, S. & Zerva, A. (2012). Very dense seismic array observations in Furukawa district, Japan, *Seismol. Res. Lett.*, 83(5), 765-774.

# CHAPTER 4

# CONVERGENCE IN UPM: APPLICATION TO VISUALIZING DATA SATURATION

## 4.1 Introduction

In this chapter, I investigate how the UPM map resolutions change as the number of observation data increase. As the number of observation data increase, the data uncertainty decreases and the reliability in the estimation of the mean increases. And if UPM really projects data uncertainty onto map resolutions, this change in data uncertainty with increasing observations must be reflected in the spatial resolutions. As we will see in this chapter, UPM map resolutions approaches that of conventional map resolutions with increasing number of observations. I utilize this information-dependent characteristic of UPM maps to address the issue of data sufficiency in spatial maps.

In the past few decades, advancement in data collection, e.g. high-resolution remote sensing, monitoring sensor networks, etc., considerably increased the availability of spatial data [**1**]. 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 tried to continue if the amount of data is determined to be not enough. This concept is based on the idea that the goal of the hazard maps is to plot accurate information using the enough data, which can contribute to the disaster reduction. On the other hand, in practice, we need to plot the maps based on only the available data. Under the situation, we have two essential questions; (1) how we judge whether amount of the available data is sufficient, and (2) how we draw the maps consistent with the data accumulation.

In this chapter, I address these issues by utilizing the information-dependent characteristic of UPM maps. I introduce KL divergence increments, based on information theory [**2**], that measures the incremental information gain as an UPM map is updated. Data saturation or sufficiency is reached when no more incremental information gain is observed even after adding new data to an UPM map.

In literature, papers addressing the issue of data saturation in spatial mapping is rare to find. There are some papers addressing the issue of optimal sampling, however, the focus is on assessing the grid layout rather than the optimal amount of data [**3,4,5**]. Past research in other fields, has seen some papers addressing the issue of data saturation **[6,7,8]**. However, many of them are qualitative in nature and

none of them considers data uncertainty in their formulations and hence the question is reliability is left unanswered.

In this study, I examine a methodology to visualize and quantify the excess or deficiency of data in mapping earthquake ground motion amplifications. In the next section, I introduce the incremental KL Divergence parameter to quantify data saturation. In section 4.3, I introduce two numerical experiments to discuss how the parameter can help in quantifying and visualizing data saturation in spatial maps. In section 4.4, I 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.

## 4.2 Methodology

### 4.2.1 Update UPM maps at multiple stages of data accumulation

In chapter 3, model evaluation in UPM was done using a *k*-fold cross-validation [**9**]. 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, cross-validation is replaced with computationally faster Watanabe-Akaike information criterion (WAIC) [**10,11**]. 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.

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 $(\mu_j)$ and standard deviation $(\sigma_j)$ depends on the amount of data. As we will see in the next sections, the UPM maps approach conventional mapping as the data increase. Quantifying this convergence process will help us find at which stage data saturation occurs.

### 4.2.2 $\Delta D_{KL}$ : Proposed parameter to quantify data saturation

The estimated mean $(\mu_j)$ and estimated standard deviation $(\sigma_j)$ improves as more and more data are added to the mapping. In other words, maps evolve with the addition of data in time. In this study, I visualize and quantify this property of maps and decide the point of data saturation, which means that more increase in data adds no more information to the map.

As we will see in section 4.3 and section 4.4, UPM approaches conventional mapping (ordinary Kriging) as the number of observation data increase. To quantify this convergence in UPM, I use a parameter based on Kullback-Leibler (KL) Divergence [**2**]. KL Divergence measures how different two probabilistic distributions are. It is usually defined as

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{4.1}$$

where $P$ and $Q$ are continuous random variables and $p$ and $q$ are the associated probability densities.

In this study, I define a quantity called incremental KL Divergence ($\Delta D_{KL}$ ) given by

$$\Delta D_{KL_{[N+\Delta N]}} = \sum_j D_{KL}\left(P_{[N],j} || P_{[N+\Delta N],j}\right) \tag{4.2}$$

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. Such convergence measure has been proposed to evaluate the performance of numerical analysis [**12**].

The parameter $\Delta D_{KL}$ measures the information gain as the maps are updated with more and more data in time. Data saturation happens when $\Delta D_{KL}$ approaches zero, which means that no more spatial information is added even upon adding more data to the map. The uniqueness of the parameter $\Delta D_{KL}$ 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 $\Delta D_{KL}$ 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 ($\mu_j$) and standard deviation $(\sigma_j)$, separately. (2) Unlike UPM, no statistical dependences between the mean and standard deviation are incorporated in ordinary Kriging. (3) Unlike ordinary Kriging, the UPM maps vary with the amount of observation data. When observation data is less, UPM maps have a low spatial resolution. The ordinary Kriging maps, on the other hand, have a high spatial resolution even when the observation data is less. 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.

## 4.3 Numerical experiments

In this section, two different numerical experiments are introduced where I investigate how the UPM map resolutions change as the number of observation data increase. In numerical experiment A, I use the same sparse data set layout as used in the previous chapter (numerical experiment B in chapter 3). In numerical experiment B, I make a hypothetical model of the wave amplification in and around an alluvial valley.

### 4.3.1 Numerical experiment A

#### 4.3.1.1 Data

**Fig. 4.1** introduces the data used for this numerical experiment. At first, random samples (shown as black dots) following normal distribution are artificially generated as observations at 50 sites in one-

dimension. The given mean values ($\mu$) of the normal distribution vary in sinusoidal manner along the site locations. The given standard deviations ($\sigma$) increase from the left to right. Then, observations from 72% of the original sites are randomly removed and sparsely distributed dataset (shown as red triangles) is created.

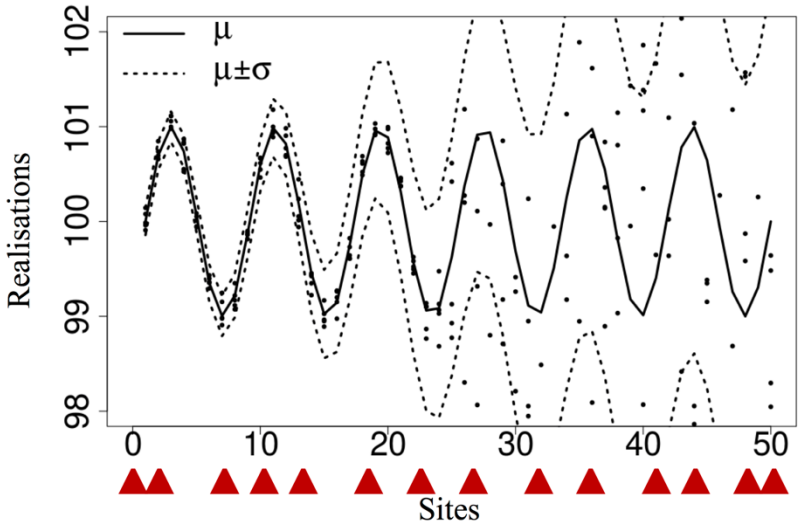



**Figure 4.1** Dataset for numerical experiment A

#### 4.3.1.2 Results

**Fig. 4.2** shows 9 different datasets with 4, 8, 16, 32, 64, 128, 256, 512 and 1024 random samples per site. Conventional mapping is shown by a black line and UPM is shown by a red line.

When the number of observations ($N$) is low, the UPM map resolution is high (follows the conventional mapping) at the low uncertainty area on the left, and the UPM map resolution is low (becomes smooth) at the high uncertainty area on the right. This effect enhances the detailed visual (mapping) on significantly different observations expected between the sites with low standard deviation and rough visual (mapping) on insignificant observations between the sites. As the number of observations ($N$) increase, UPM is observed to no longer possess the smooth nature at high uncertainty areas and starts approaching the conventional mapping. This change in the characteristics of the UPM with the increase in number of observations has a significance in understanding the population.

When the number of observations is low, there is less information for modeling and so, the estimated model parameters are quite unstable. The conventional mapping for low observation dataset when compared with the known mean values, is erroneous. The error increases as the uncertainty increases on the right. In such a situation, the smoothness introduced in the UPM on the higher uncertainty side is a better representative of the data than the erroneous conventional mapping. However, when the number of observations is high, there is more information for modeling and so, the estimated model parameters are stable. The conventional mapping for 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 observation increases.

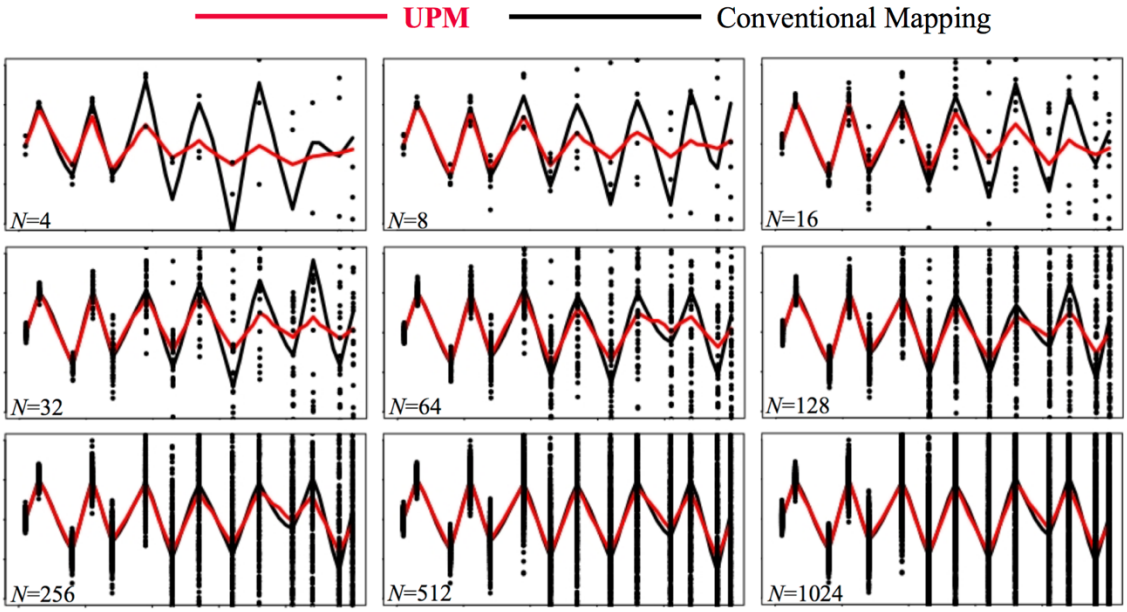



**Figure 4.2** Evolution of UPM maps compared with conventional mapping in numerical experiment A

**Fig. 4.3** shows the incremental KL divergence ($\Delta D_{KL}$) with respect to the number of observations, calculated using **equation (4.2)**. Sites located at the edges are not included in the calculation of $\Delta D_{KL}$. This is because I would like to discuss the results as interpolation problem. At the edge, the values are estimated as extrapolation problem. It is observed that $\Delta D_{KL}$ 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 evolution of $\Delta D_{KL}$ through the UPM.

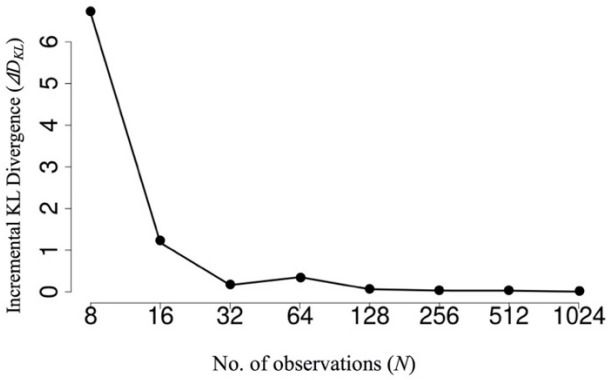



**Figure 4.3** Plot of $\Delta D_{KL}$ vs $N$ for the UPM maps in numerical experiment A

### 4.3.2 Numerical experiment B

#### 4.3.2.1 Data

In engineering seismology, it is well known that alluvial deposits can significantly affect the amplitudes of the incident seismic waves [**13**]. It is well established with high level of confidence (low uncertainty) that the waves get 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 rock site and alluvial valley. This is because the incident angle and frequency contents are well affected.

For the numerical experiment B, I make a hypothetical model of the wave amplification in and around an alluvial valley as a spatial process (**Fig. 4.4**). 63 sites are considered in one dimension. The sites are all equally spaced except at the boundary between rock site and alluvial where the sites are more densely spaced. The reason behind this is to properly model the change in uncertainty as one move away from high uncertainty at the boundary to the low uncertainty regions.

Random wave amplification samples (mean values shown as black dots in **Fig. 4.4**) are artificially generated as observations at all these sites. Each observation sample refers to wave amplification observed for an incident seismic wave. The random samples at each site follow a lognormal distribution.

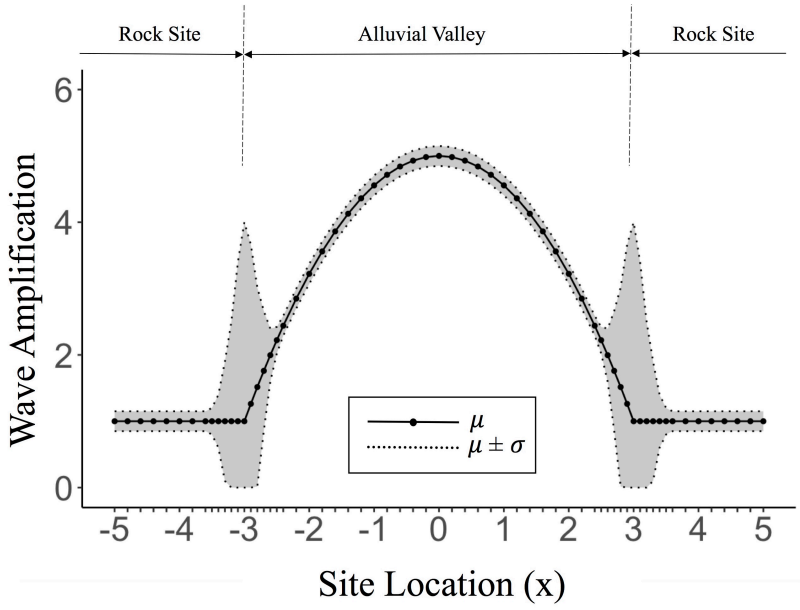



**Figure 4.4** Dataset for numerical experiment B

The known arithmetic means ($\mu$) increase as a parabolic curve from the rock site to the center of the valley representing high wave amplification. The standard deviation ($\sigma$) is treated as the uncertainty, which is high at the boundary zone of the rock site and the alluvial valley and low otherwise. The extreme amplitude changes at the basin boundary occurs as the incident angle and frequency contents are well affected at the basin boundary [**13**].

#### 4.3.2.2 Results

**Fig. 4.5(a)** shows eight different cases with 8, 16, 32, 64, 128, 256, 512 and 1024 artificial random samples per sites. 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 low number of observations ($N$), the UPM shows a smooth transition at the highly uncertain boundary zone between rock site and alluvial valley, unlike the Kriging map which is very rough and fluctuating. The boundary zone has a high $\sigma_j$. So, UPM makes the transition smooth by imposing a low $s_j$ in the boundary zone. However, in the low uncertainty areas including the center of the valley, UPM behaves like Kriging. UPM keeps the Kriging shape in areas of low $\sigma_j$ by imposing a high $s_j$ around $j$. As the number of observations ($N$) increase, UPM starts to approach the Kriging map, as also observed in numerical experiment A.

**Fig. 4.5(b)** shows the incremental KL divergence ($\Delta D_{KL}$) with respect to the number of observations. It is observed that $\Delta D_{KL}$ 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.

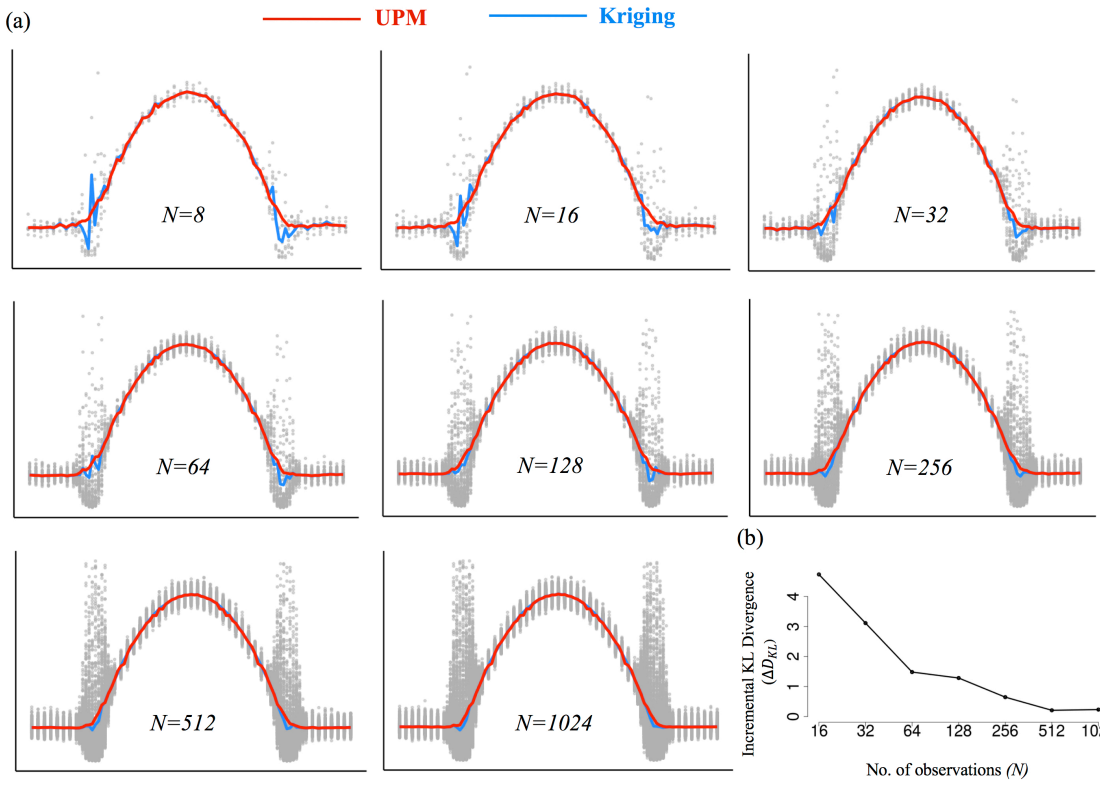



**Figure 4.5** (a) Evolution of UPM and Kriging maps for numerical experiment B (b) Plot of $\Delta D_{KL}$ vs $N$ for the UPM maps

## 4.4 Application: A case study in Furukawa, Japan

### 4.4.1 Data

In this case study, earthquake data collected over 7 years from 31 sites in the very dense seismic array of Furukawa, Japan [**14**] is used to generate an amplification map in the area. 176 earthquake events recorded between 29th October, 2011 and 19th September, 2018, were used for the analysis. **Fig. 4.6** shows the layout of the seismic array consisting of 31 seismometers in the significantly damaged area in Osaki city.

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. The average peak ground acceleration (PGA) in the recorded events ranges from 6 gal to 119 gal.

The availability of observation data varies with the site locations. For studying 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.

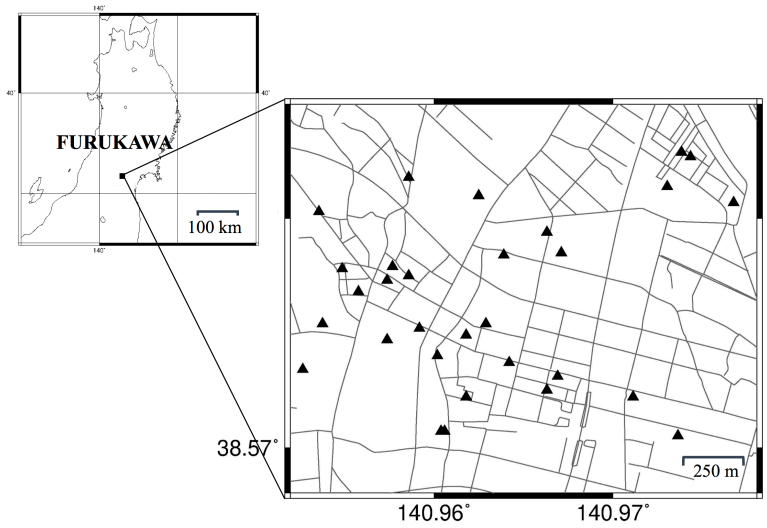



**Figure 4.6** Spatial distribution of seismometers (▲) in Furukawa district, Japan

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) and peak ground velocity (PGV) at site *j* to the spatial average calculated over all the available sites during a one earthquake event. The PGA and PGV are calculated from the vector sum of EW component and NS component of the earthquake record. To generate a UPM map of the site amplification, the dataset comprised of 431 sites with 31 measurement sites from the seismic network and 400 missing sites, all distributed in a rectangular grid.

### 4.4.2 Results

**Figs 4.7(a) and (b)** show the site amplification maps calculated using PGAs. For all the datasets, UPM has been compared with Kriging maps. It is observed that 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 increase, the two maps start becoming more and more similar. If we focus on how the UPM changes with the increase in the number of observations, it is observed that as more and more earthquake data are included, spatial variation starts appearing on the map and starts to converge as the number of events increase. To discuss this convergence quantitatively, **Fig. 4.7(c)** shows a plot of $\Delta D_{KL}$ with the *N*, the number of observations. The $\Delta D_{KL}$ is calculated only for the available sites common to all the events. It is shown that as the number of observations increase, the $\Delta D_{KL}$ 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.

**Figs 4.8(a) and (b)** shows 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 to be smoother in comparison to the PGA plots. As observed in the case of the PGA plots, in this case also the UPM lots start to converge with the increasing number of the observations. **Fig. 4.8(c)** 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 a 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.

## 4.5 Discussion

It is evident from the cases discussed in sections 4.3 and 4.4, that the optimum number of data which is deemed enough to extract useful information depends on available dataset. In the case of numerical experiments, data saturation is attained after 128 observations (numerical experiment A) and 512 observations (numerical experiment B) have been collected. However, in the case of the seismic array in Furukawa, Japan, 176 observations seem to be enough to understand the population statistics.

The reason for this difference might be explained based on the record to record variability present at the sites. For example, in the numerical experiment B, the peak of true record to record variability was high ($\sigma_j$=3) at the boundary zone. Thus, more data is necessary to accurately estimate the mean and the record to record variability (standard deviation) at 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.

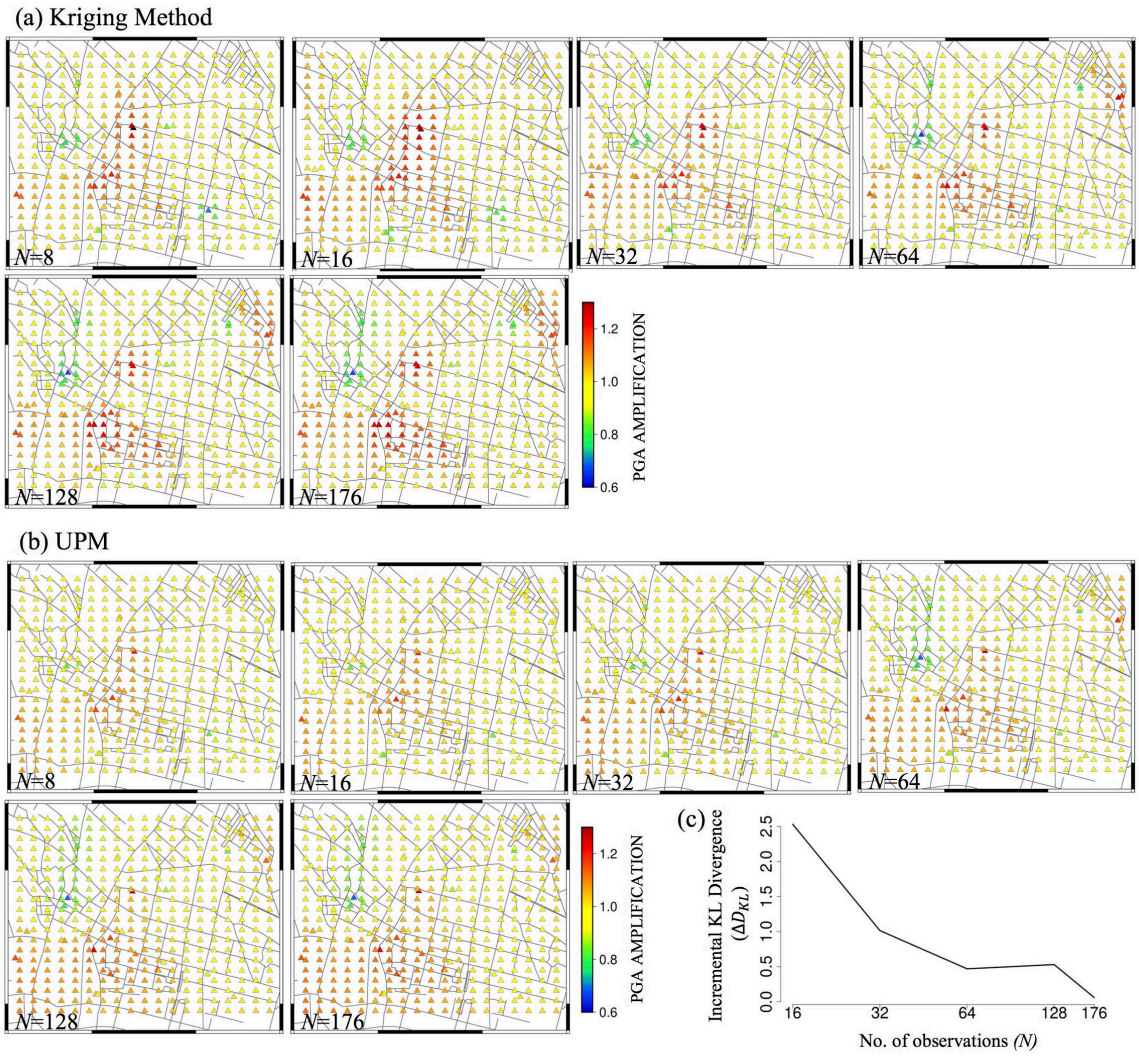



**Figure 4.7** (a) Evolution of Kriging maps of PGA amplifications in Furukawa district, Japan (b) Evolution of UPM maps of PGA amplifications in Furukawa district, Japan (c) Plot of $\Delta D_{KL}$ vs $N$ for the UPM maps of PGA amplifications

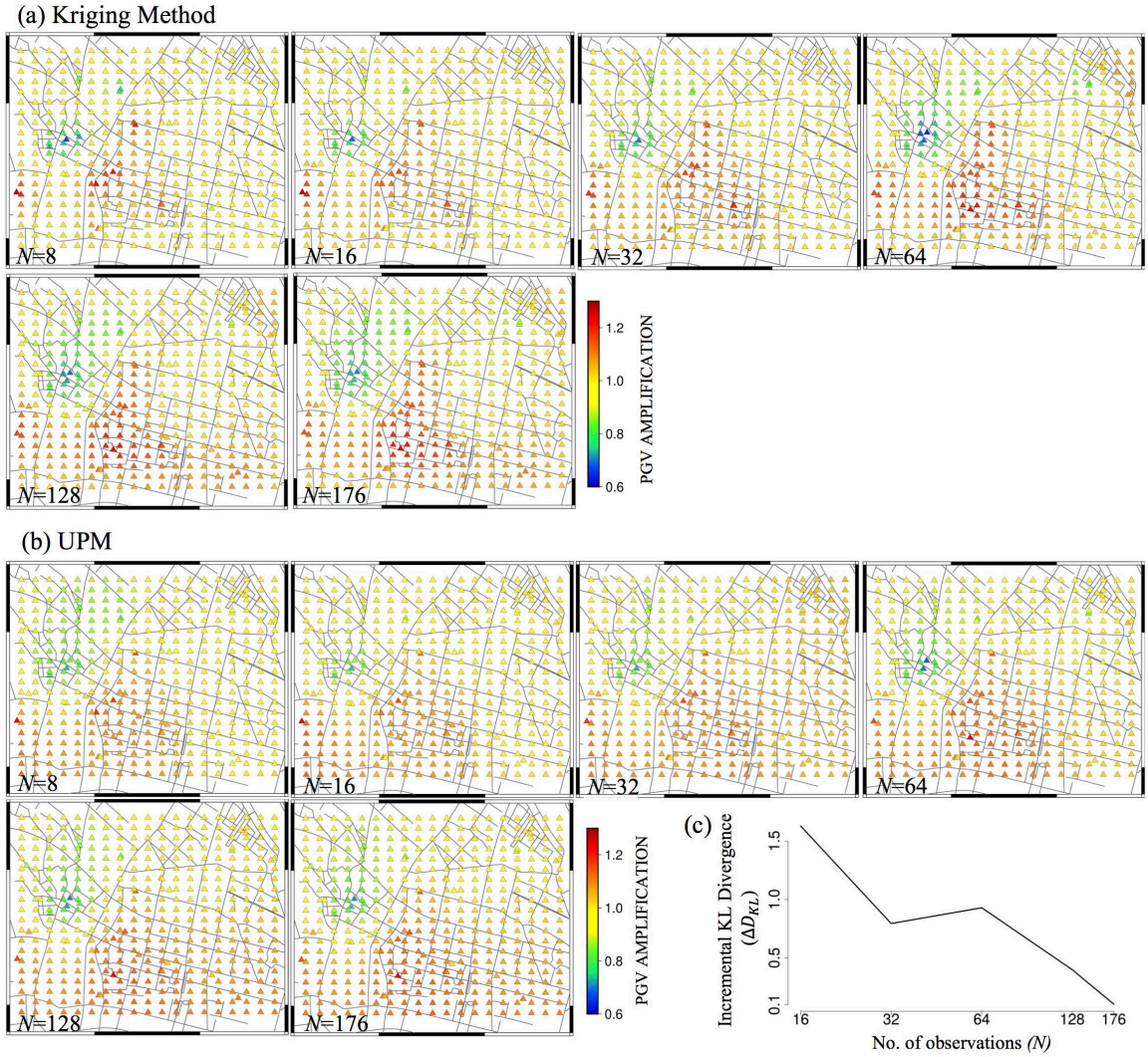



**Figure 4.8** (a) Evolution of Kriging maps of PGV amplifications in Furukawa district, Japan (b) Evolution of UPM maps of PGV amplifications in Furukawa district, Japan (c) Plot of $\Delta D_{KL}$ vs $N$ for the UPM maps of PGV amplifications

For the case study area in Furukawa, Japan, I used 400 missing sites to create the PGA and PGV site amplification maps from 31 measurement sites, and the convergence process is quantified based on the maps obtained. However, the convergence process won't 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 process between the PGA and PGV site amplification maps are not identical. One reason could be that the PGA and PGV process are not the same. The spatial distribution patterns in **Fig. 4.7** and **Fig. 4.8** are clearly different. This means that the spatial datasets of PGA and PGV are different. Another reason could be the difference in information gain process for the two. Unless two processes have the same incremental information gain and same record to record variabilities at all locations, it will be rare for two processes having the same convergence process based on information theory.

## 4.6 Conclusion

In this chapter, I investigate how the UPM map resolutions change as the number of observation data increase and confirm if UPM really projects data uncertainty onto map resolutions. We observe that UPM map resolutions approaches that of conventional map resolutions with increasing number of observations. I utilize this information-dependent characteristic of UPM maps to address the issue of data sufficiency in spatial maps.
As a measure of data saturation, I define a parameter $\Delta D_{KL}$, based on information theory, which quantifies information gain as maps are updated with new data over time. Data saturation happens when $\Delta D_{KL}$ approaches zero, which means the no more spatial information is getting added to the maps and we can stop updating the maps.
The numerical experiment results showed 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 observation increases.
The concept is then applied to a case study area in Furukawa district of Japan where earthquake data is 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 a data saturation and from the viewpoint of information theory, the current operation may be terminated, and the seismometers may be rearranged to resolve map in the unclear areas.

## Data availability

The seismic array data from Furukawa, Japan used in the application (Section 4.4) 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 which are in the significantly damaged area are utilized. The seismometers not considered in this study are F15, F21, F30, F32, F36 and F37.

## References


1. Lee, J. G., & Kang, M. (2015). Geospatial big data: challenges and opportunities. Big Data Research, 2(2), 74-81.
2. Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79-86.
3. Hughes, J. P., & Lettenmaier, D. P. (1981). Data requirements for kriging: estimation and network design. *Water Resources Research*, *17*(6), 1641-1650.
4. Wang, J. F., Stein, A., Gao, B. B., & Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics*, *2*, 1-14.
5. James, B. R., & Gorelick, S. M. (1994). When enough is enough: The worth of monitoring data in aquifer remediation design. *Water Resources Research*, *30*(12), 3499-3513.
6. Chaudhuri, S., Motwani, R., & Narasayya, V. (1998). Random sampling for histogram construction: How much is enough?. *ACM SIGMOD Record*, *27*(2), 436-447.
7. Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field methods*, *18*(1), 59-82.
8. Fusch, P. I., & Ness, L. R. (2015). Are we there yet? Data saturation in qualitative research. *The qualitative report*, *20*(9), 1408.
9. Stone, M. (1974). Cross-validation and multinomial prediction. *Biometrika*, *61*(3), 509-515.
10. Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, 11, 3571-3594.
11. Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, *24*(6), 997-1016.
12. Goto, H., & Bielak, J. (2008). Galerkin boundary integral equation method for spontaneous rupture propagation problems: SH-case. *Geophysical Journal International*, *172*(3), 1083-1103.
13. 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.

14. Goto, H., Morikawa, H., Inatani, M., Ogura, Y., Tokue, S., Zhang, X.R., Iwasaki, M., Araki, M., Sawada, S. & Zerva, A. (2012). Very dense seismic array observations in Furukawa district, Japan, *Seismol. Res. Lett.*, 83(5), 765-774.

**CHAPTER 5**

**APPLICATION OF UPM: UPDATING MAP RESOLUTIONS OF A CONVENTIONAL MAP**

## 5.1 Introduction

In this chapter, I investigate how the framework of UPM can be applied to update the map resolutions of a conventional map to incorporate the data uncertainty from local sources. As an example of conventional map, I focus on the J-SHIS map of site amplification factor [**1**]. And as local data, I use soil boring data provided by Geo-Research Institute (GRI) in the case study area.

The site amplification factor map is based primarily on the engineering geomorphic classification map which offers the geomorphic classification in a standard area mesh in whole of Japan, approximately with a spatial resolution of 250 m [**2**]. A 250 m-mesh is a square area of 7.5 arc-seconds latitude and 11.25 arc-seconds longitude (about 250 m × 250 m) [**3,4**]. Wakamatsu and Matsuoka [**4**] developed a linear regression formula for individual engineering geomorphic units to estimate the associated average shear velocity in the upper 30 m depth of soil (AVs30) using elevation, slope and distance from mountain as the explanatory variables. Using this linear regression formula, the map of AVs30 is prepared at first [**5**]. The site amplification factor map is then calculated from the map of AVs30 using Fujimoto and Midorikawa [**6**]. The site amplification factor means amplified ratio calculated from the engineering bedrock (Vs=400 m/s) up to the ground surface. **Figs 5.1, 5.2 and 5.3** show the J-SHIS map of engineering geomorphic classification, AVs30 and site amplification factor, respectively. However, local soil data is not incorporated in the mapping process.

In this chapter, the objective is to update the resolutions of the J-SHIS map of site amplification factor to incorporate the data uncertainty from local sources. **Fig. 5.4** shows the location of Ibaraki - Takatsuki area of Osaka, the case study area. **Fig. 5.5** shows the J-SHIS map of site amplification factor in the case study area. **Fig. 5.6** shows the distribution of soil boring locations in the case study area.

In technical terms, Bayesian update of the J-SHIS map of site amplification factor is summarized by **equations (5.1)** and **(5.1)**. Using the fundamental idea of *Bayes' rule* as introduced in chapter 2, we can write

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)} \tag{5.1}$$

Where A refers to amplification factor at site and D refers to data from boring.

$$P(A|D,J) = \frac{P(J|A)P(A|D)}{P(J)} \tag{5.2}$$

Where J refers to the J-SHIS values.

Thus, $P(A|D,J)$ is the updated amplification at site. The existing J-SHIS site amplification factors are used as a prior information and the site amplification factor calculated from available boring data is used as observation data (or likelihood information).

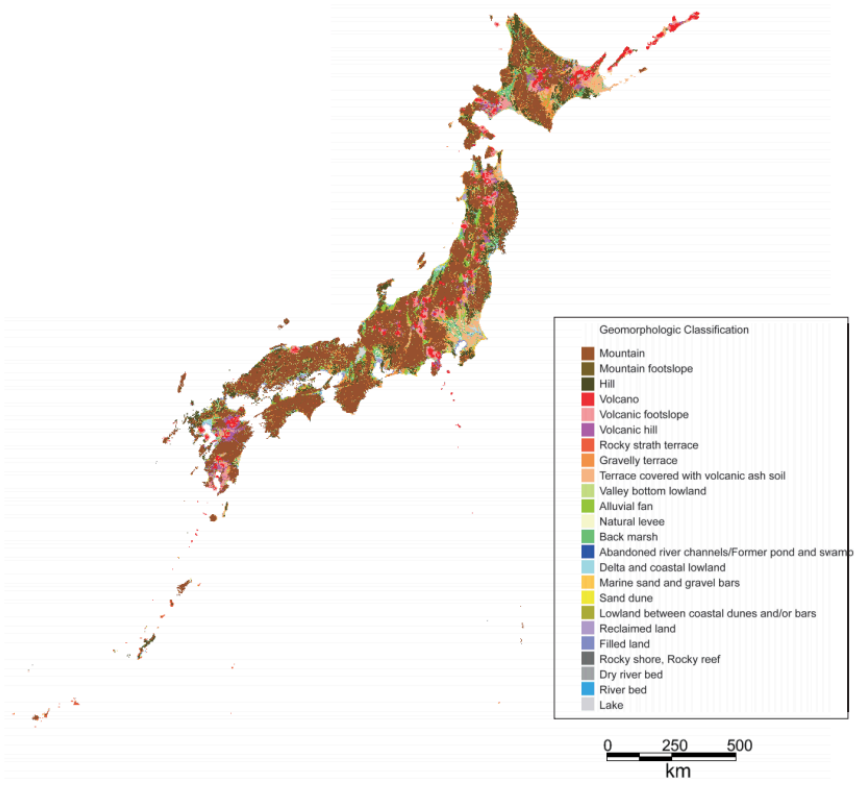



**Figure 5.1** 250 m-mesh J-SHIS map of engineering geomorphic classification **[4]**

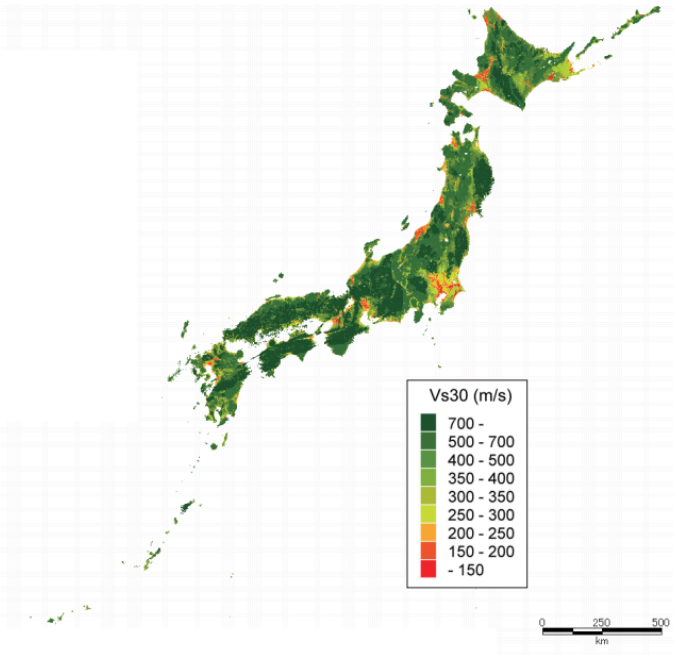



**Figure 5.2** J-SHIS AVs30 map calculated from the engineering geomorphic classification map **[4]**

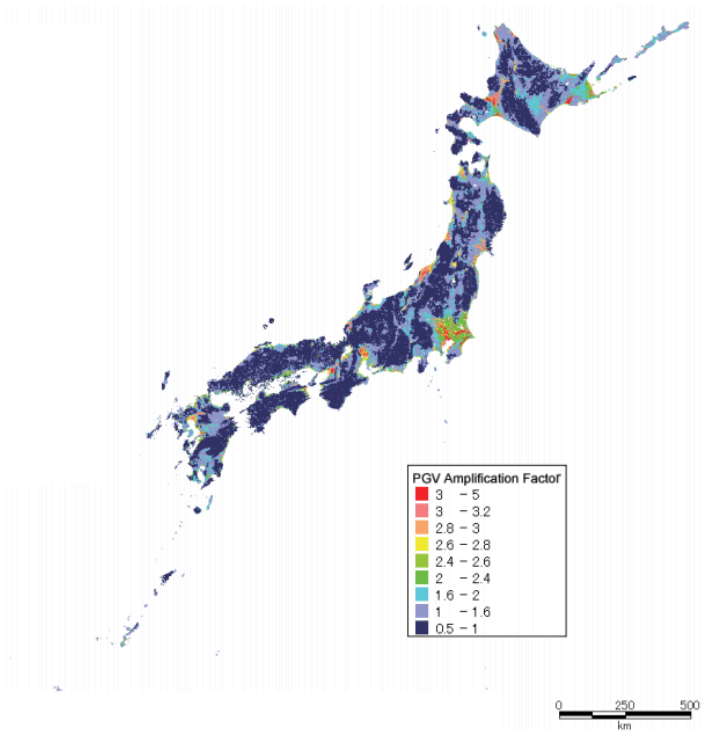



**Figure 5.3** J-SHIS map of PGV site amplification factor converted from the AVs30 map **[4]**

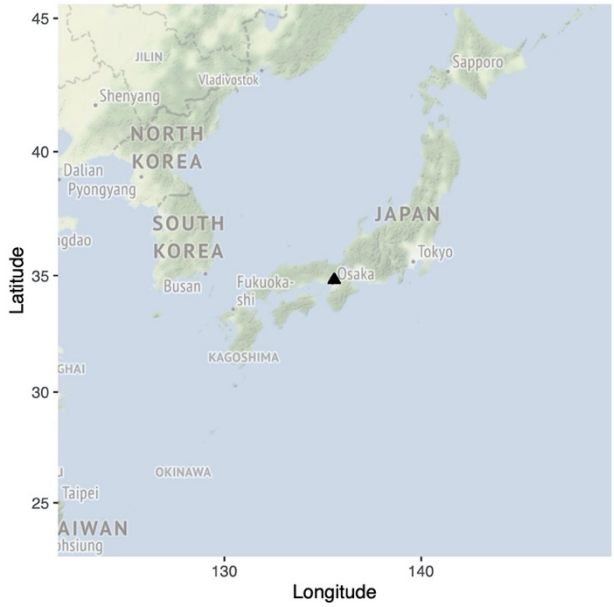



**Figure 5.4** Location of case study area in the map of Japan

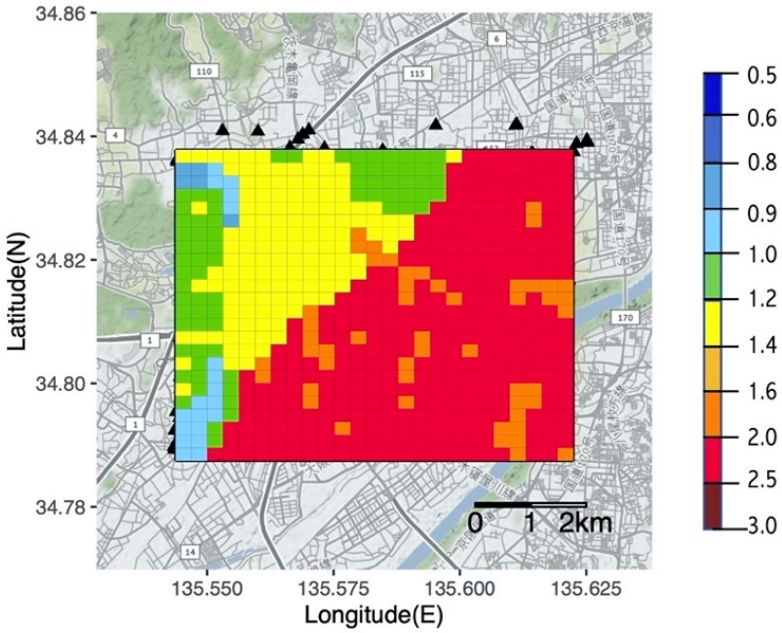



**Figure 5.5** J-SHIS map of site amplification factor in the case study area

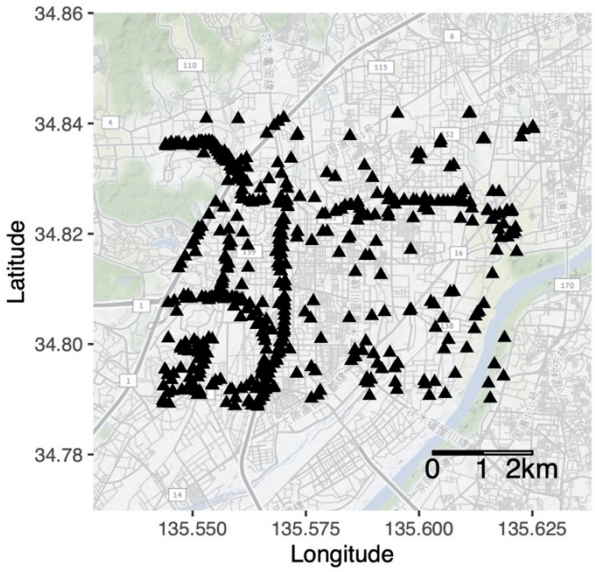



**Figure 5.6** Distribution of soil boring locations in the case study area

I use informative priors for the mean and standard deviation of amplification values at each site. For the mean of the prior, I use the existing site amplification factor value at individual meshes. As for the standard deviation of the prior, I use the available information in literature related to the uncertainty for individual geomorphic units and regression equation of AVs30 and site amplification factor. Wakamatsu and Matsuoka [**4**] obtained standard deviation values for individual geomorphic units by a regression analysis. Similarly, Fujimoto and Midorikawa [**6**] also obtained a standard deviation value while establishing a relationship between AVs30 and site amplification factor. I use a vector sum of both these standard deviations to create the standard deviation of the prior.

$$\sigma_p = \sqrt{(\sigma_1^2 + \sigma_2^2)} \tag{5.3}$$

Where $\sigma_p$ refers to the standard deviation of prior, $\sigma_1$ refers to the standard deviation from Wakamatsu and Matsuoka [**4**] and $\sigma_2$ refers to the standard deviation from Fujimoto and Midorikawa [**6**]. **Table 5.6** shows the calculated $\sigma_p$ for the engineering geomorphic units in the case study area.

In sections 5.2 and 5.3, the seismic ground response analysis and the concept of engineering seismic base layer are introduced as they are crucial to the calculation of site amplification factors. In section 5.4, the soil structure in the case study area is investigated in detail. In section 5.5, the process of calculating site amplification factor is explained. In sections 5.6 and 5.7, the observation data for the UPM is introduced and the resulting updated J-SHIS map of site amplification factor is compared with its original counterpart.

**Table 5.1** $\sigma_p$ for the engineering geomorphic units in the case study area.

| Engineering Geomorphic Unit | $\sigma_1$ | $\sigma_2$ | $\sigma_p$ |
|---|---|---|---|
| Gravelly terrace | 0.122 | 0.166 | 0.206009709 |
| Hill | 0.175 | 0.166 | 0.24120738 |
| Valley bottom lowland | 0.158 | 0.166 | 0.229172424 |
| Alluvial fan | 0.116 | 0.166 | 0.202514197 |
| Natural levee | 0.124 | 0.166 | 0.207200386 |
| Back marsh | 0.116 | 0.166 | 0.202514197 |
| Dry riverbed | - | 0.166 | - |

## 5.2 Seismic ground response analysis

In this study, seismic ground response analysis is employed to calculate the local site amplification factors from the soil boring data. As soil shows non-linear behavior at very small strains, a simple model such as bilinear model which are applicable to structures, is not applicable to soil. Computer programs for seismic ground response analysis needs to consider the stress-strain relationships and soil particle-pore water mixture [**7**]. The general concept used for calculating the cyclic shear deformation characteristics is the hysteric stress-strain curve expressed by (secant) shear modulus and damping ratio as a function of shear strain amplitude [**8,9,10**].

SHAKE was the first computer program that aimed for analyzing the behavior of ground during earthquakes[**11**]. It solved the equation of motion in a frequency domain by employing a Fourier series expansion and the concept of complex modulus. The latter is necessary to consider nonlinear behavior in a linear system. This technique is popularly known as equivalent linear method. SHAKE became popular because of its easy handling characteristics and is still used with minor modifications. However, although SHAKE popularized the equivalent linear method, it is still an approximate method. It has a few limitations and many attempts were made to overcome it. Improvements were made by considering frequency-dependent characteristics in stiffness and/or damping.

In this study, however, a linear model of soil was used, and seismic ground response analysis is performed based on the fundamental multiple reflection theory.

## 5.3 Engineering seismic base layer

In Japan, during seismic ground response analysis, input earthquake motions are defined at the engineering seismic base layer [**7**]. Engineering seismic base layer is required to set the boundary conditions of the equation of motion, so that the behavior outside the boundary does not affect the result of the seismic ground response analysis.

Previously, seismic bedrock was used as the definition of the engineering seismic base layer [**12,13**]. However, with improvement in research, the depth of the seismic bedrock kept on increasing. In practice,

it is nearly impossible to make investigations at such deep depths. Thus, for engineering concerns, the definition of seismic bedrock and engineering seismic base layer needs to be separate. Also, there is plenty of observed earthquake motion records at the engineering seismic base layer as compared to the deep seismic bedrock, making it logical to define the earthquake motion at this depth [**14**].

The definition of engineering base layer in the design specifications of Japan is usually defined by the S-wave velocity. The engineering seismic base layer defined for port facilities [**15**] and for road bridge [**16**] correspond to the layers with SPT *N*-value of 50. In the definition used by buildings [**17**], layers deeper than SPT *N*=50 layer is used as the engineering seismic base layer as a layer with SPT *N*=50 appears when Vs is a little larger than 300 m/s in many cases.

In this study, I define sand or gravel layers with Vs ≥ 400m/s as the engineering seismic base layer. A considerable thickness and presence of local continuity has also been considered.

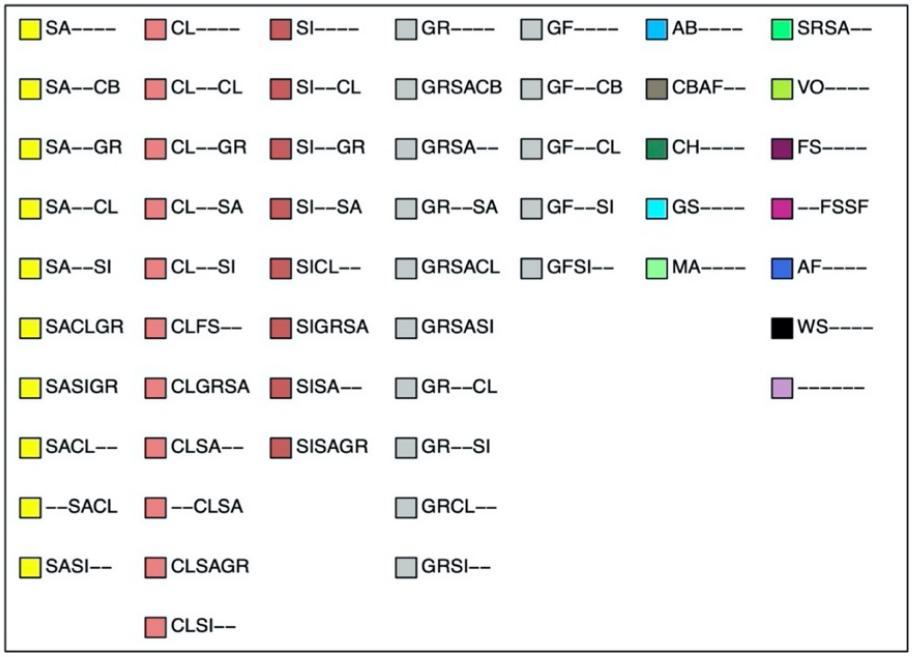
## 5.4 Investigation of soil structure in the case study area

### 5.4.1 Boring data

In this section, I investigate the soil structure in and around Ibaraki-Takatsuki area of Osaka. For this investigation, I use boring data from 3779 locations. The data was provided by Geo-research institute (GRI). The different soil types present in the case study area and the associated color legend used in this study are summarized in **Fig. 5.7**.

In **Fig. 5.8,** the case study area is shown by the red rectangle. A and B are two main sections having a high density of boring data. At first, the general soil structure in the case study area is discussed using these two main sections. Section A, to some extent, follows the decrease in basin elevation from the mountain to the river. **Fig. 5.9** shows a close up of section A. The cross-section of soil structure along section A (**Fig. 5.10**) shows a gravel layer that continues almost across the whole section. Above the gravel layer, there are sediments whose depth increase from the left to the right and reach a maximum depth near the riverbed of Yodo river. In the middle of the section, the sediments are clay dominated, however, as we move closer to the riverbed the sand percentage increases. In order to confirm if we can use this gravel layer as the engineering seismic base layer, I superimpose the zones with blocks of SPT-*N* ≥ 50 in **Fig. 5.11**. It is observed that although the gravel layer in the right side (points 4~6) can be considered to be the engineering seismic base layer, however, the gravel layer on the left side cannot be treated as the engineering seismic base layer as it barely reaches SPT-*N* ≥ 50. Also, in the middle of the section A (near point 4) data is not enough to confirm if it is the same gravel layer that connects from left to right. Thus, based on the data along section A, it seems that a unique engineering seismic base layer is difficult to assign for the case study area.

| | | | |
|---|---|---|---|
| SA | Sand（砂） | AB | Ballast（バラス） |
| CL | Clay（粘土） | AF | Asphalt（アスファルト） |
| SI | Silt<br>（シルト） | CH | Chert (Sedimentary Rock)<br>（チャート） |
| GR | Gravel（礫） | GS | Rubble（捨石・ホルダー） |
| GF | Gravelly soil<br>（礫質土） | MA | Weathered Granite（まさ土） |
| CB | Cobble stone<br>（玉石） | SR | Volcanic soil (Kyushu)<br>（しらす） |
| FS | Humus Soil<br>（腐植土） | VO | Volcanic Soil<br>（火山灰土） |
| SF | Sandy Soil<br>（砂質土） | WS | Waste<br>（廃棄物） |




Example: SACLGR : SA(Main) + CL(Secondary) + GR(Minor)
SA---- : SA(Main)
SA--GR: SA(Main) + GR(Minor)

**Figure 5.7** Color legend for different soil types in the boring data

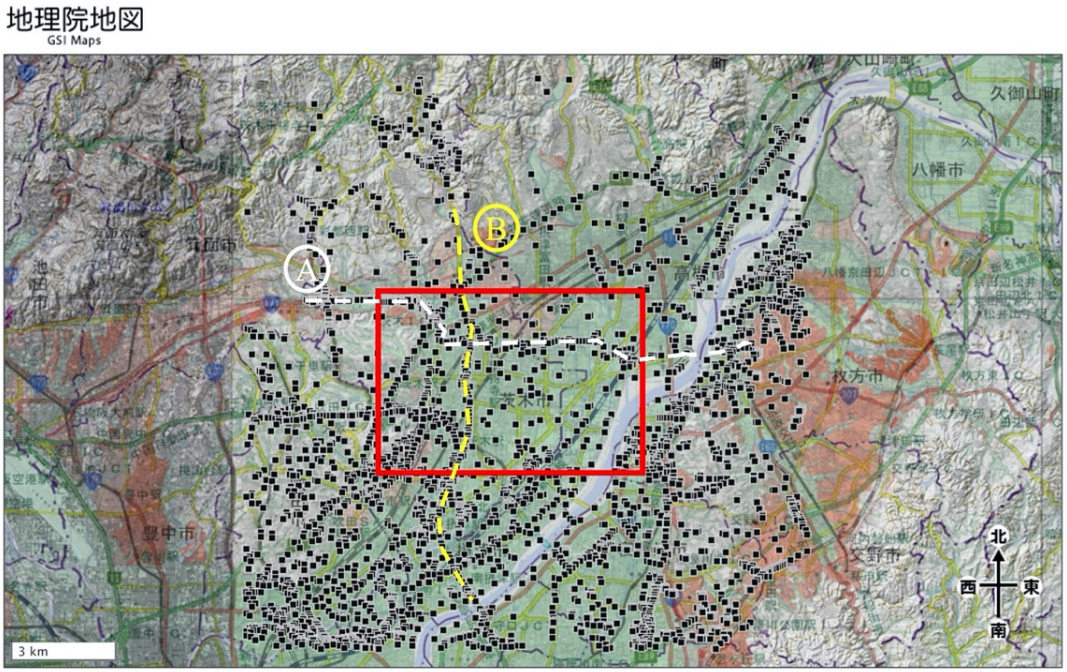



**Figure 5.8** The two main sections A and B having a high density of boring data

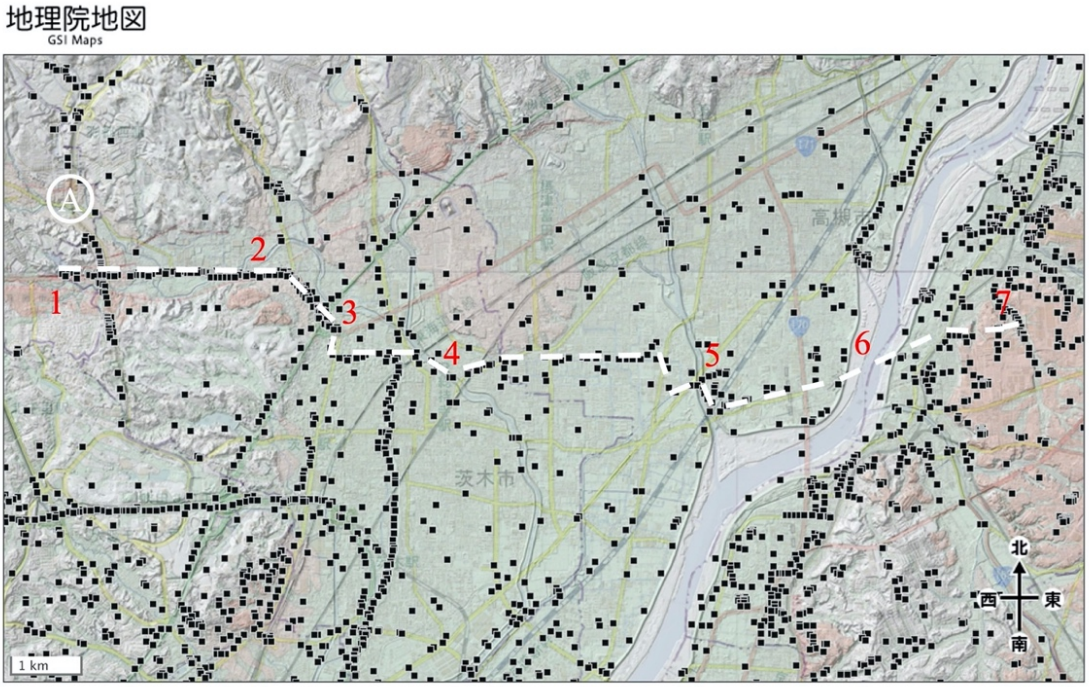



**Figure 5.9** A close-up of section A

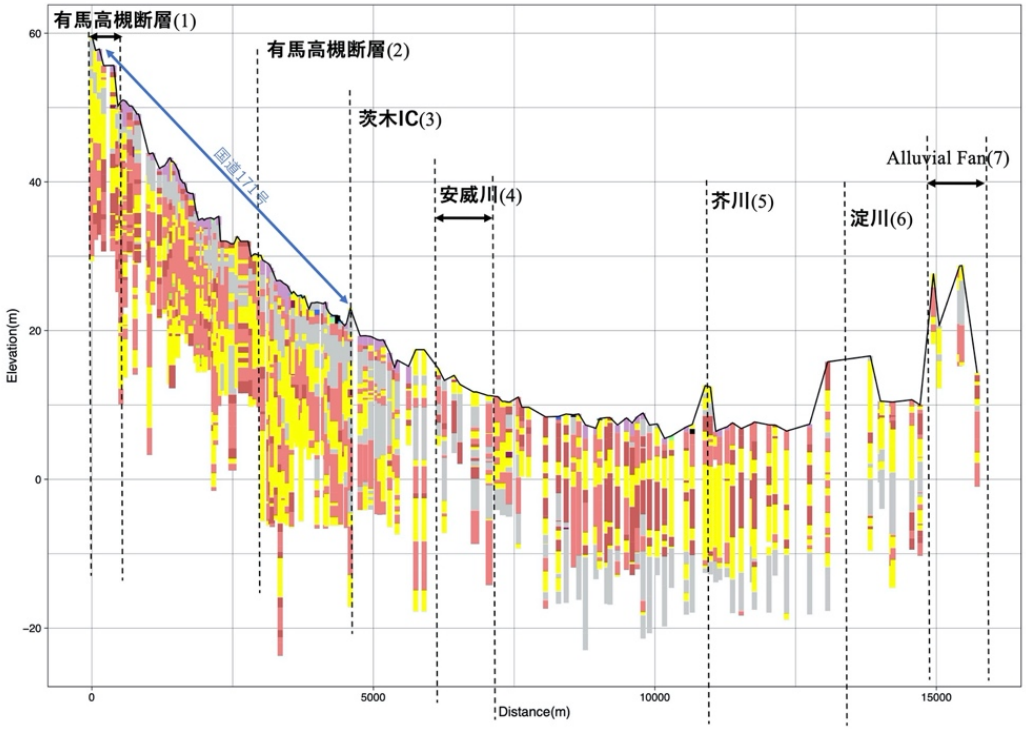



**Figure 5.10** Soil structure along section A

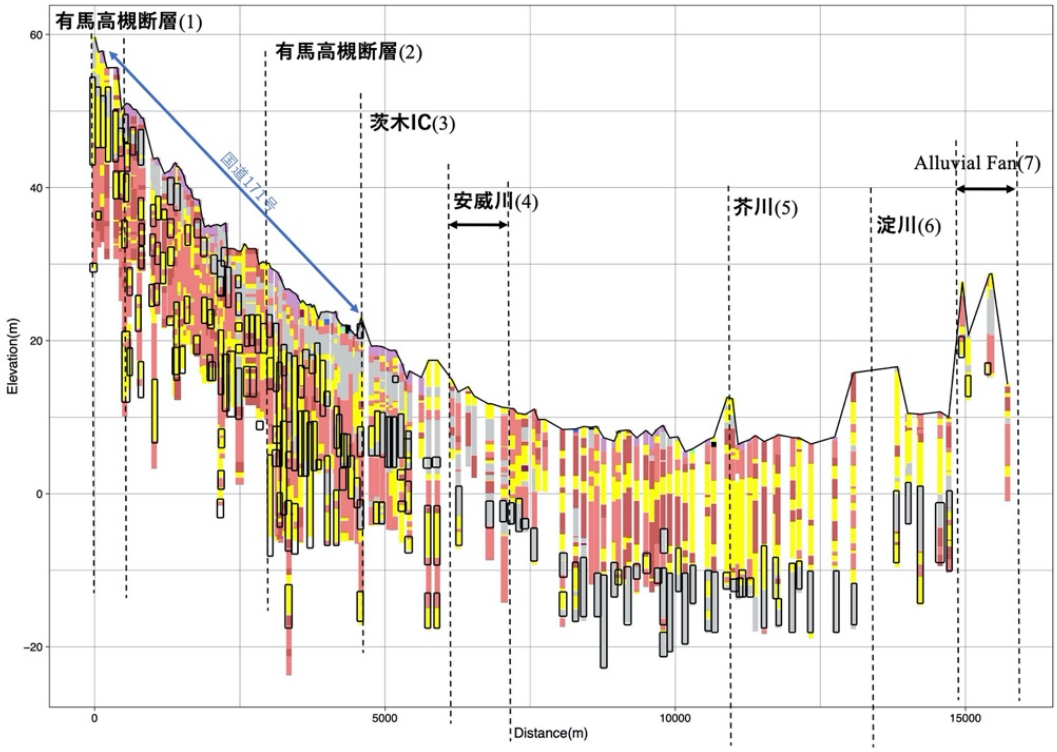



**Figure 5.11** Zones of SPT-$N \geq 50$ superimposed in the boring data along section A

Next, I investigate the soil structure along section B. **Fig. 5.12** shows a close up of section B. In section B, points 3~4 is along an old river channel, which explains the gravel layer close to the ground (**Fig. 5.13**). Superimposing the zones with SPT-$N \geq 50$, it can be seen that this gravel layer could be considered as the engineering seismic base layer (**Fig. 5.14**). However, around point 4, the available data is not enough to make a decision on the engineering seismic base layer. However, except near the Yodo river, it is not possible to assign an engineering seismic base layer. As observed in section A, here also the sand dominance increases as the section is closer to the river.

The two main sections A and B help in making a general image of the soil structure in the case study area. However, there are many portions where the data along the section was not adequate. In order to try and understand the soil structure more detail, I investigate the area further with multiple cross-section cuts.

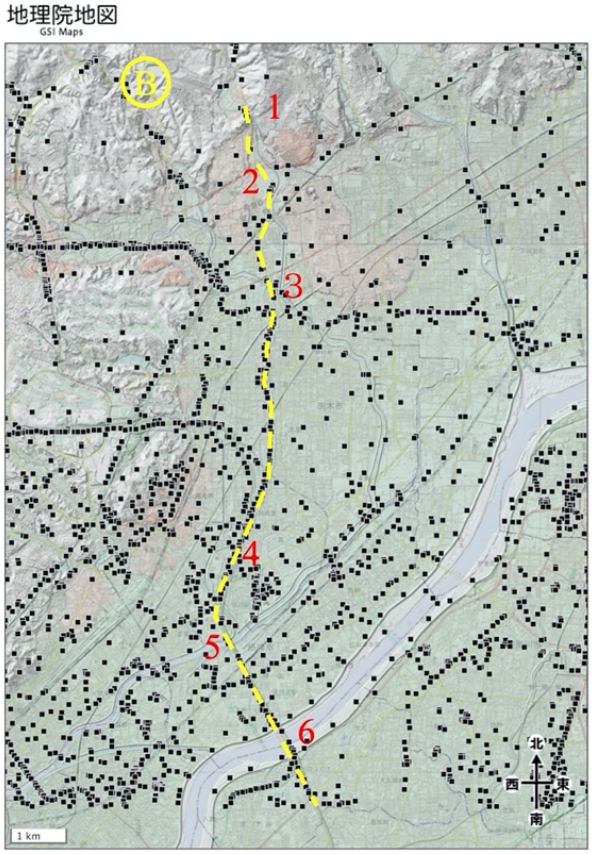



**Figure 5.12** A close-up of section B

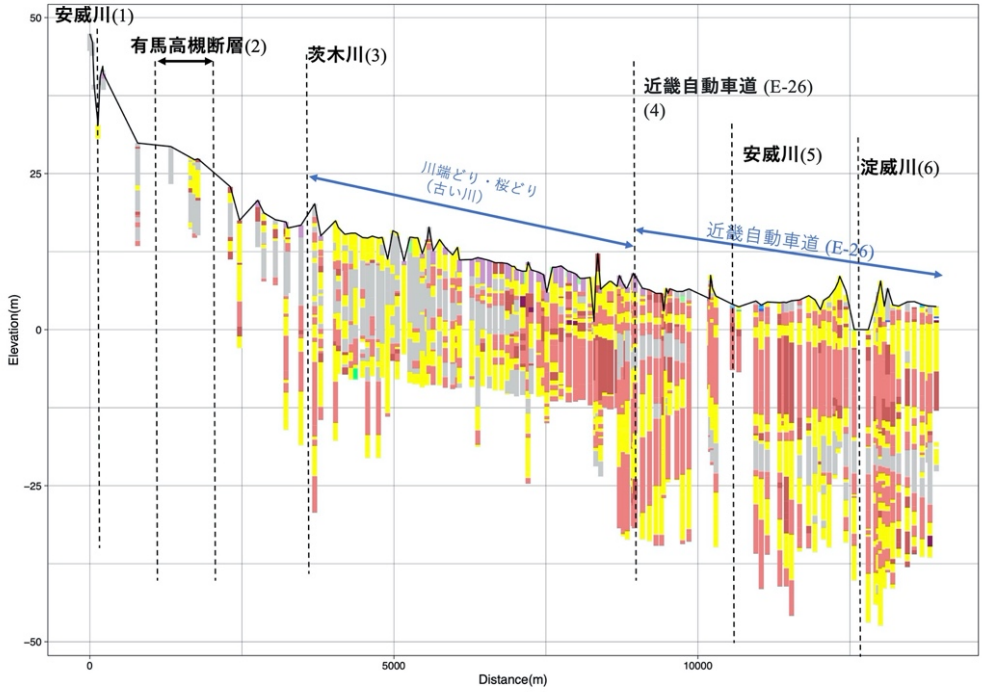



**Figure 5.13** Soil structure along section B

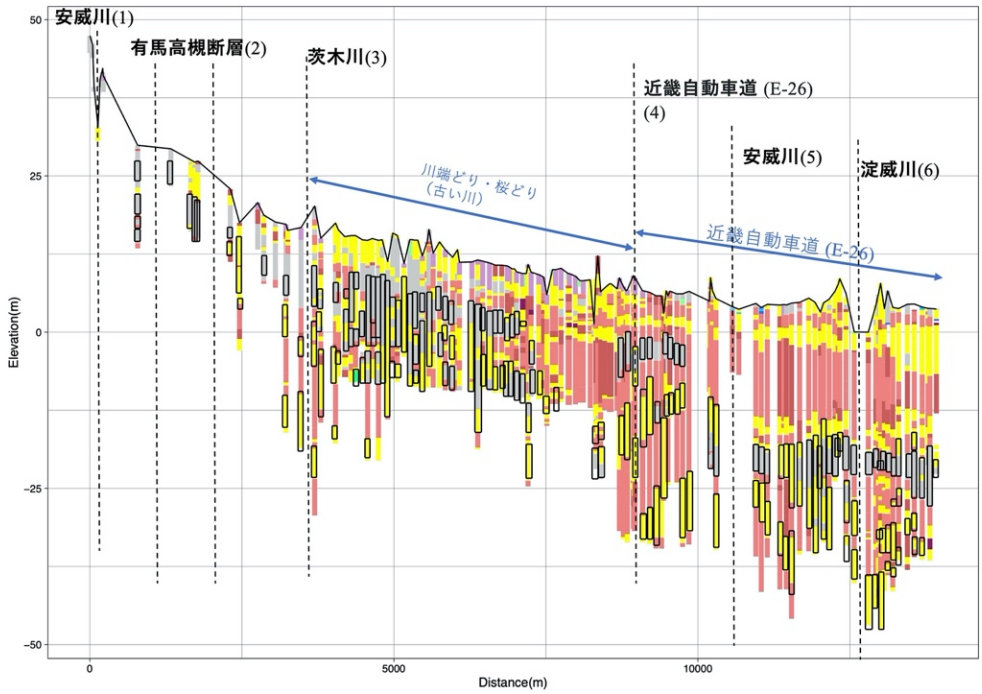



**Figure 5.14** Zones of SPT-$N \geq 50$ superimposed in the boring data along section B

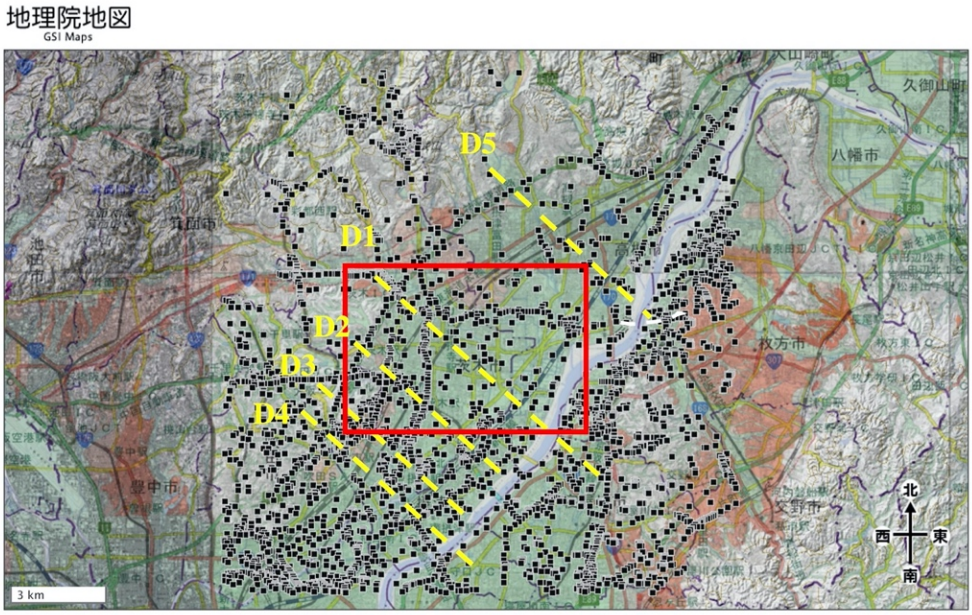



**Figure 5.15** Sections D1~D5 to investigate the soil structure in more detail

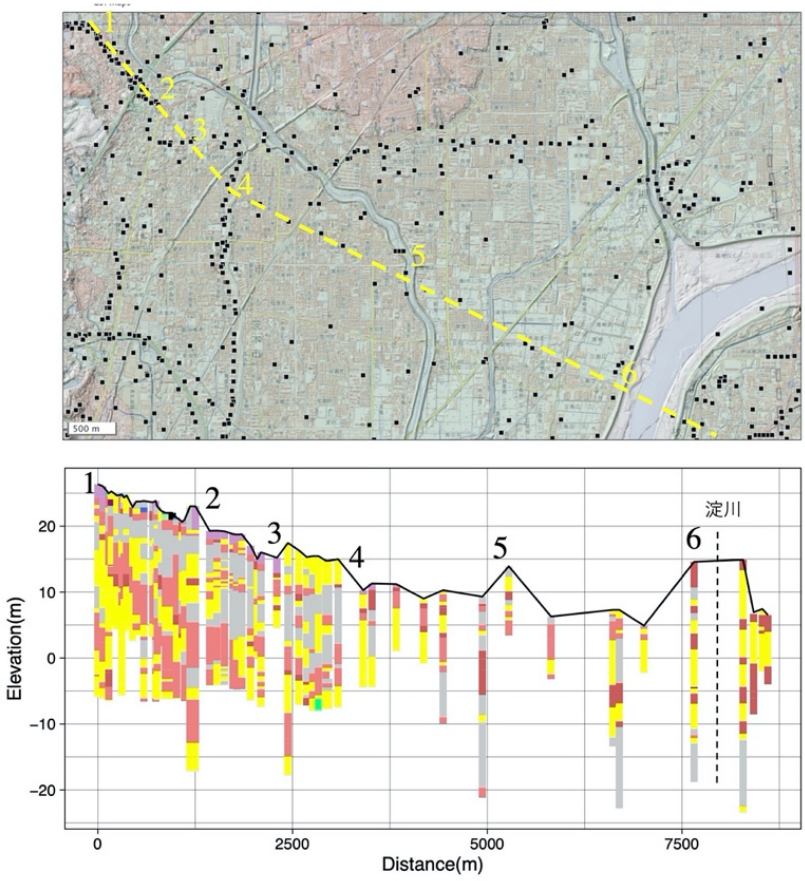



**Figure 5.16** Soil structure along section D1

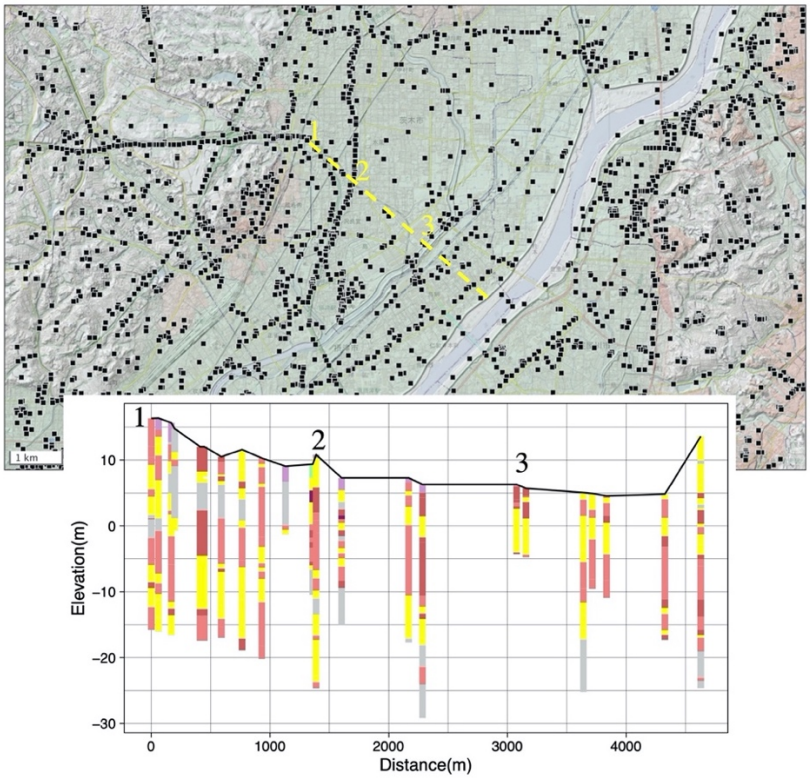



**Figure 5.17** Soil structure along section D2

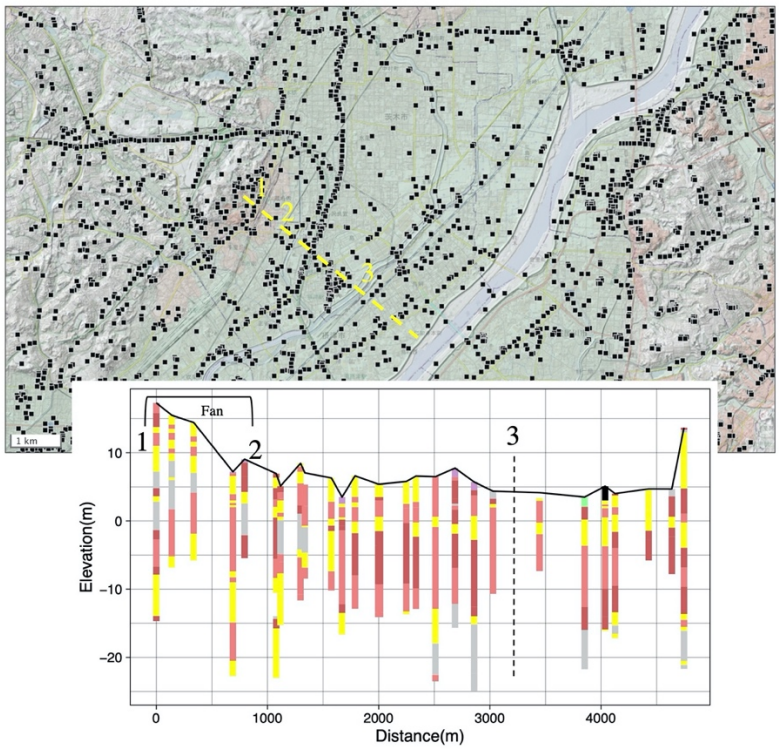



**Figure 5.18** Soil structure along section D3

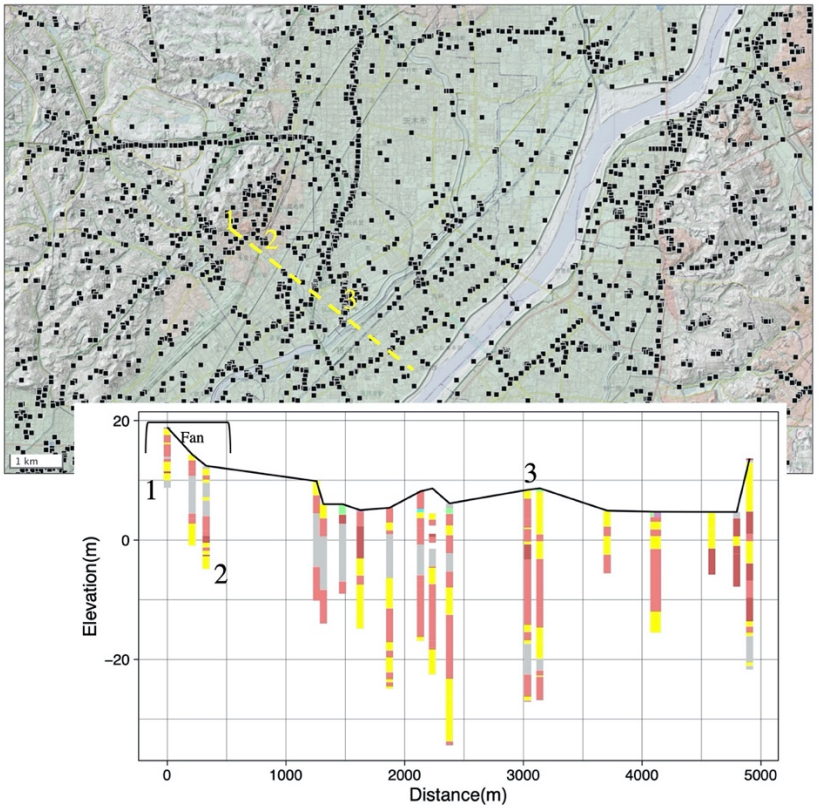



**Figure 5.19** Soil structure along section D4

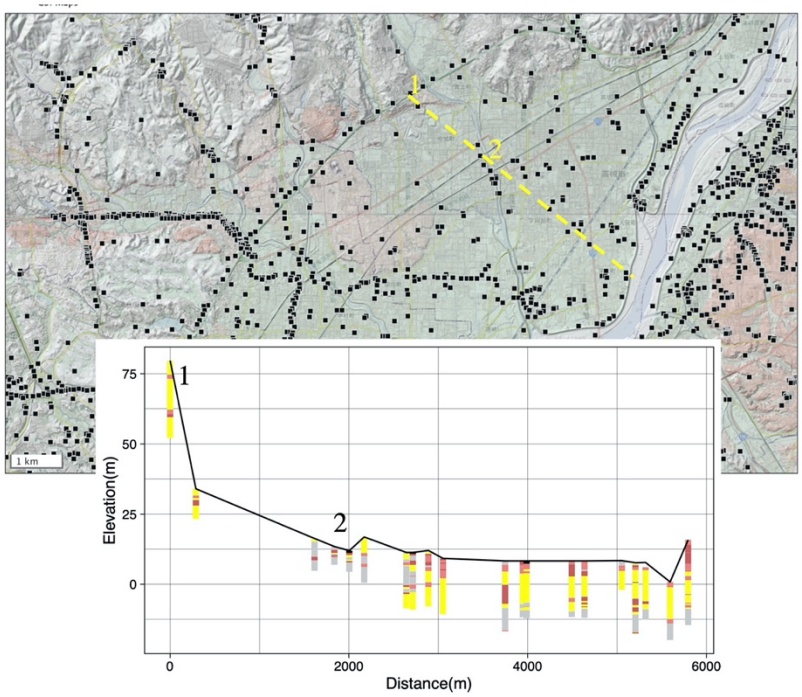



**Figure 5.20** Soil structure along section D5

**Fig. 5.15** shows the traces of sections D1~D5 to understand how uniformly the depth of sediment changes from the mountains to the riverbed of Yodo river. As can be seen from all the sections D1~D5, the maximum depth of sediments is similar near the Yodo riverbed and there is an increase from the mountains to the riverbed (**Figs 5.16 ~ 5.20**). However, as the data is sparse near the riverbed, it is difficult to understand the engineering seismic base layer. One more thing to note is the middle section in D2~D4 has deep clay deposits and the data is not sufficient in any of the sections to understand the engineering seismic base layer. More sections need to be investigated in this area to understand the distribution of the gravel layer.

**Fig. 5.21** shows a section along the Yodo river to understand the distribution of maximum sediments in the area. It is seen that the maximum depth of sediments near the riverbed is around 20~25 m. However, the sediments are more clay dominated along points 2~3, whereas along points 3~5 the sediments are more sand dominated. This information is more clearly depicted in the zoning part as discussed in the next section.

**Fig. 5.22** shows a section parallel to the river line going through the midway between the mountain and the river. The reason of investigating this section is to try and understand the deep clay deposits in that zone. However, along points 2~3 the section is not deep enough for the engineering seismic base layer to be found. Also, the boring data is pretty sparse along points 3~5 to understand how the engineering seismic base layer might be changing.

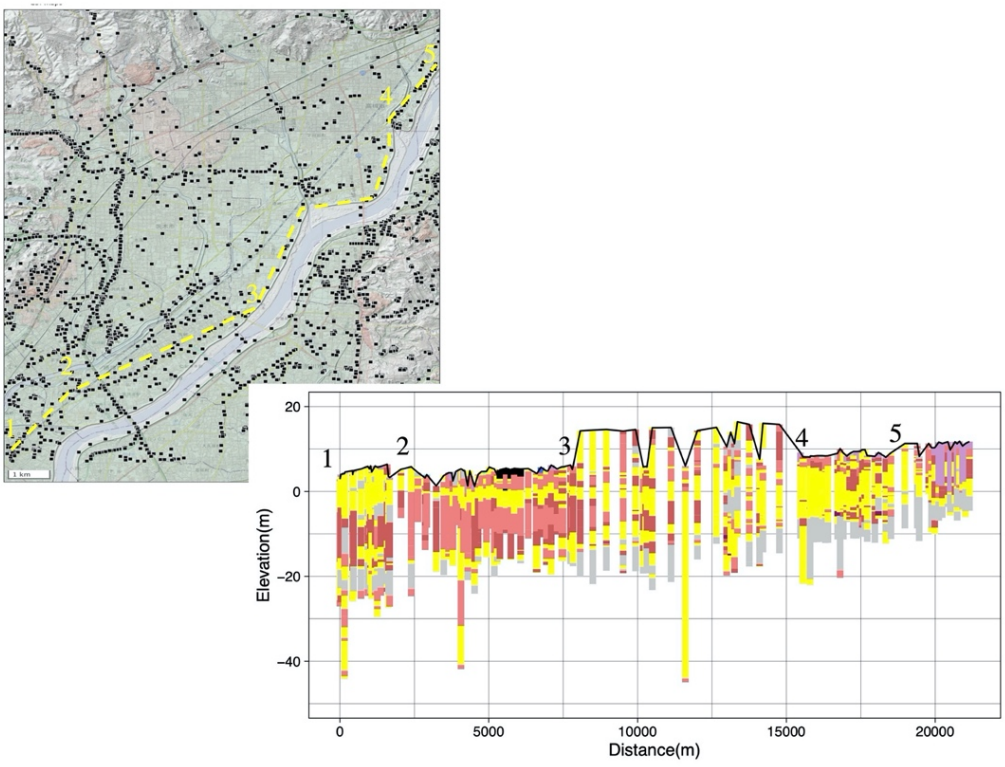



**Figure 5.21** Soil structure along Yodo river

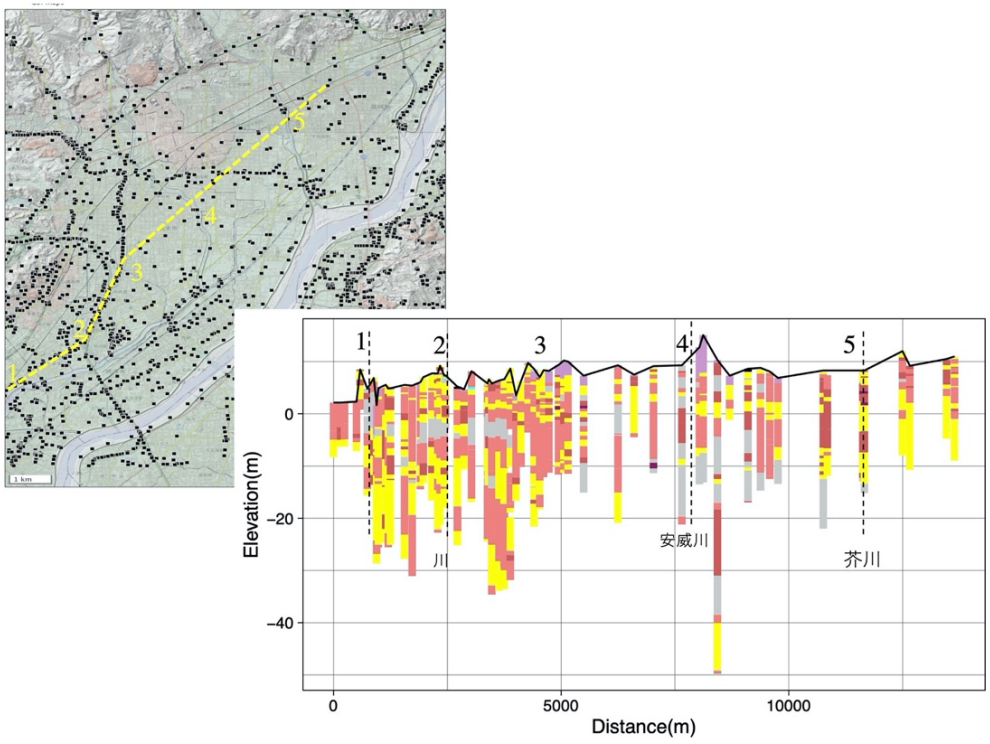



**Figure 5.22** Soil structure along a long line parallel to the river

**Figs 5.23~5.25** show multiple sections cut to understand the soil layering in the alluvial fans. However, as the data is very sparse it is not possible to conclude anything with high confidence.

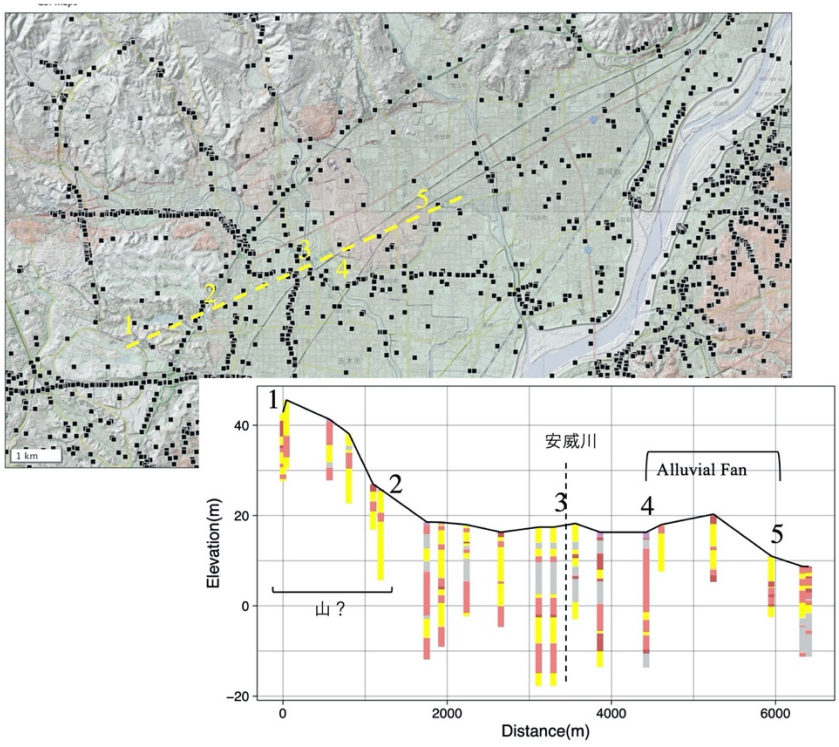



**Figure 5.23** Soil structure along section-I cutting through the alluvial fans

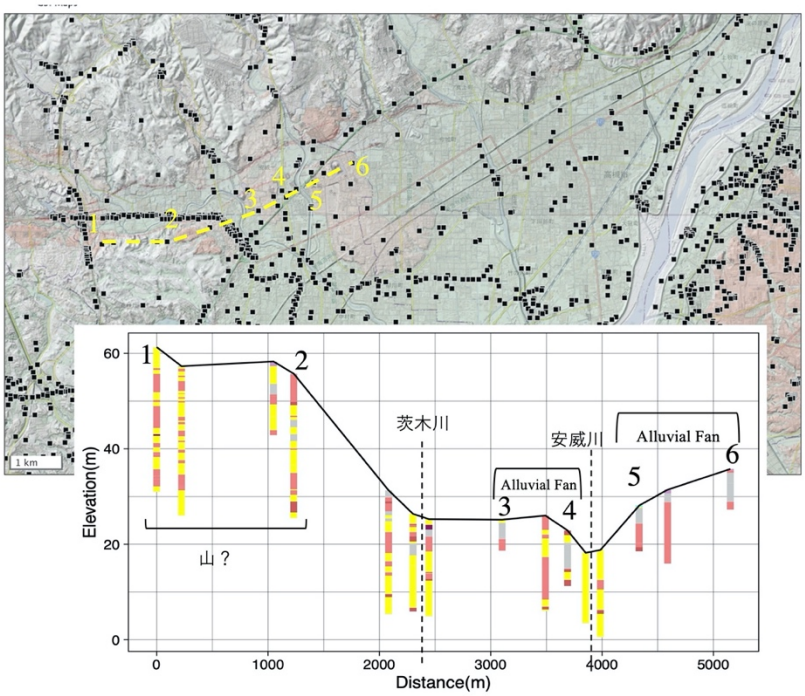



**Figure 5.24** Soil structure along section-II cutting through the alluvial fans

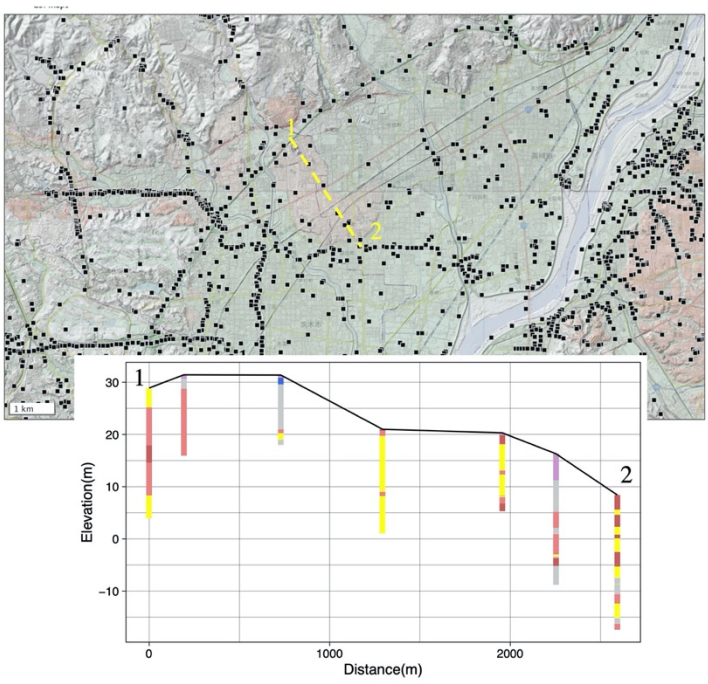



**Figure 5.25** Soil structure along section-III cutting through the alluvial fans

**Figs 5.26~5.28** show multiple sections to understand the soil layering in an old hill. In **Fig. 5.26**, section along points 1~2 cuts through a hill which is predominantly sand dominated. It is difficult to establish a unique engineering seismic base layer as the local variation is very high. In such cases, the engineering seismic base shear layer needs to be considered to be varying from point to point.

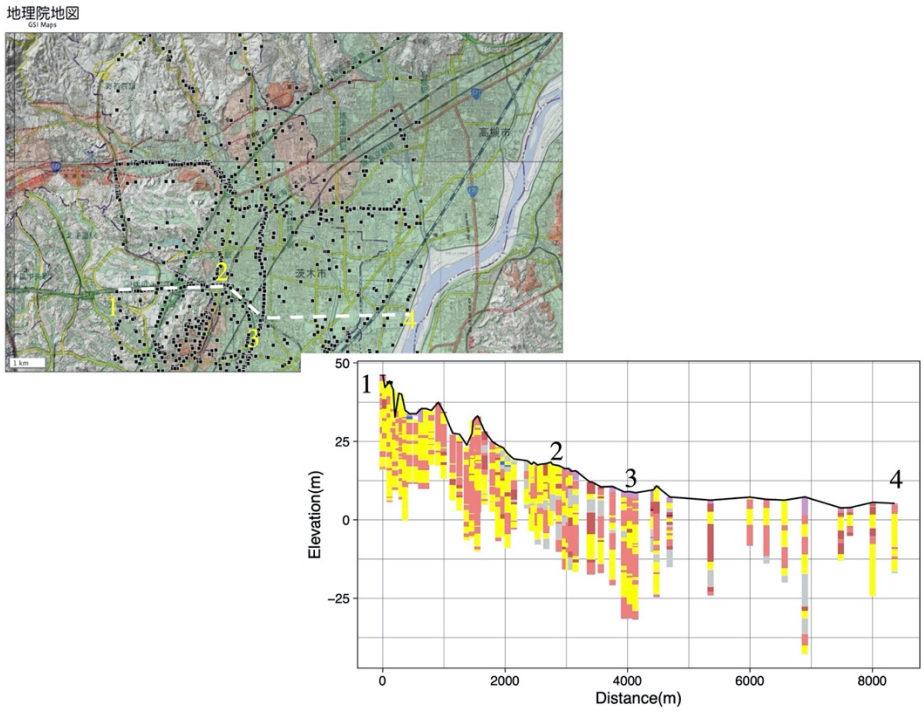



**Figure 5.26** Soil structure along section-I cutting through an old hill

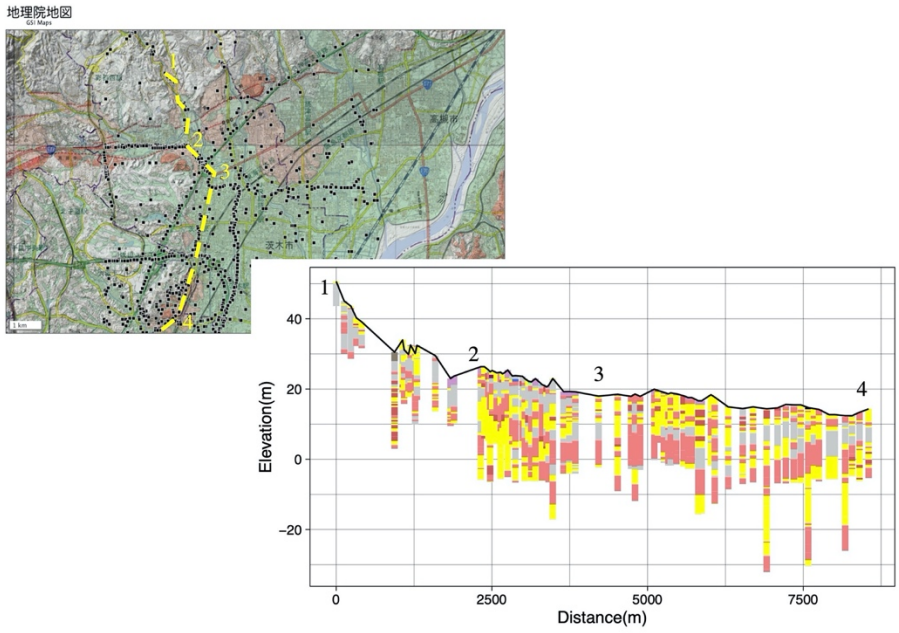



**Figure 5.27** Soil structure along section-II cutting through an old hill

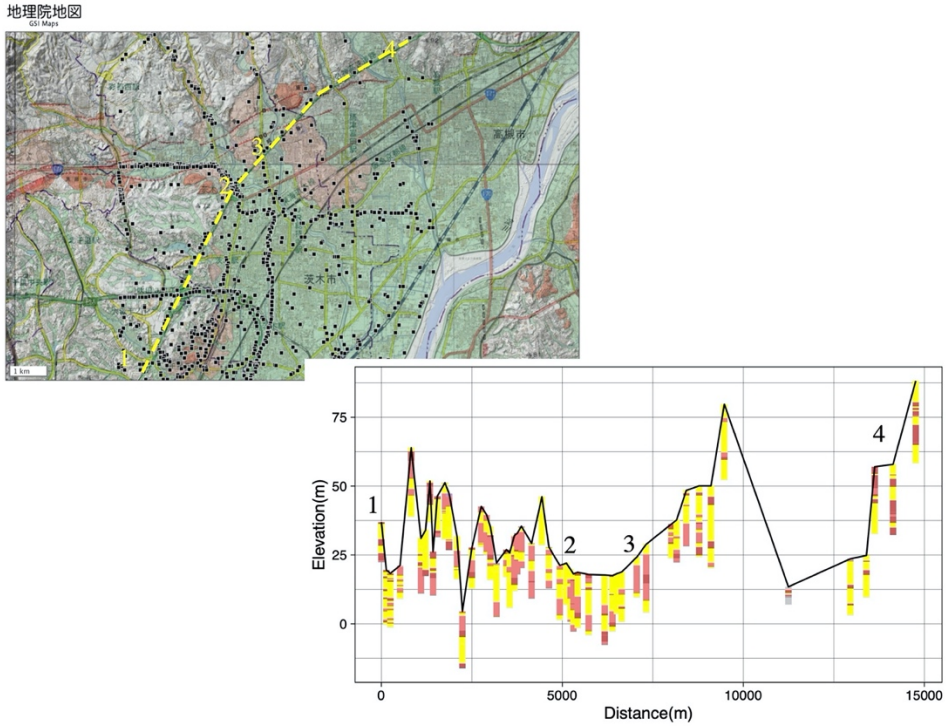



**Figure 5.28** Soil structure along section-III cutting through an old hill

### 5.4.2 Zoning

In this section, I summarize our understanding of soil structure in the case study area by defining certain zones. The zoning, as shown in **Fig. 5.29**, is based on the soil type. The yellow zone, red zone and grey zone represents the sand dominant, clay dominant and gravel dominant soil. The orange zone represents the soil where it is difficult to differentiate between sand and clay dominant. The blue zone represents those areas where the sand and clay content and the sediment thickness are highly fluctuating. The blue zone is more complicated than the orange zone and hence, is separated. The rest of the zones which doesn't fall into any of the defined zones are termed as complicated zones. In this definition of zoning, the varying thickness of sediments has also been highlighted. For very less sediments (<10 m), the zones are borderless. For thin sediment depth (10~20 m), the border is a thin line. And for thick sediment depth (~20 m) the border is marked by a thick line.

The zoning shows the presence of a distinct thick sand dominant zone in the riverbed of Yodo river. And the soil near the mountain and the alluvial fan area is mostly both sand and clay dominant. And in the intermediate zone between the mountain and the river, there exists a thick clay dominant zone. The gravel dominant zones are pretty limited and is found near some alluvial fans and close to the mountains.

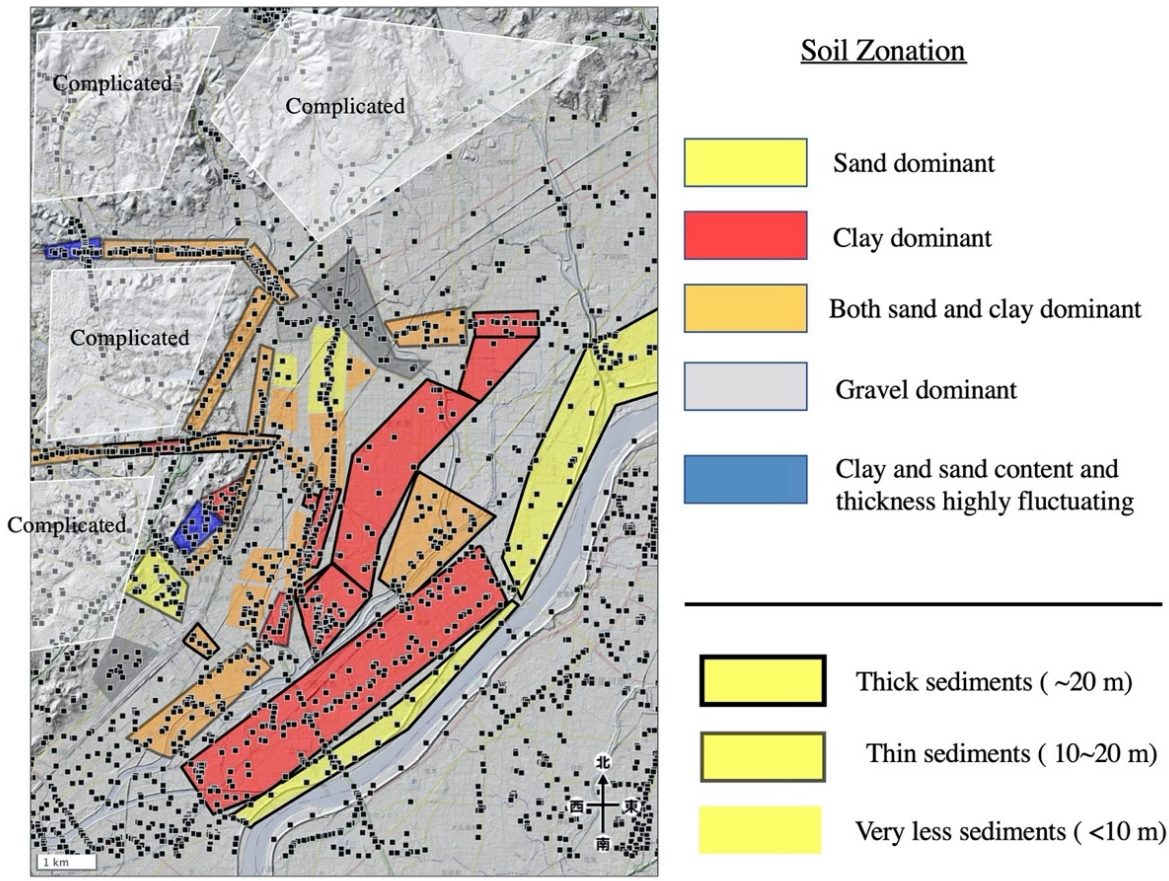



**Figure 5.29** Zoning based on soil type

As discussed, the zoning is primarily based on the soil type. However, an important information while doing seismic ground response analysis is the knowledge of the engineering seismic base layer. **Fig. 5.30** uses the zoning based on soil type to highlight the presence or absence of a unique engineering seismic base layer. In areas where a unique engineering seismic base layer can be defined, the depth of sediments won't vary significantly and can be modelled uniformly. However, in areas a unique engineering seismic base layer cannot be defined, the depth of the engineering seismic base layer and thus, the depth of the sediments will have a high local variation. In this case study area, only a few zones are observed to have a unique engineering seismic base layer. **Fig. 5.30** shows those areas with a white border. C refers to clay dominant, S refers to sand dominant and CS refers to both sand and clay dominant. In these zones, it is assumed that the engineering seismic base layer can be represented by a unique layer and thus can be modelled with the same representative section.

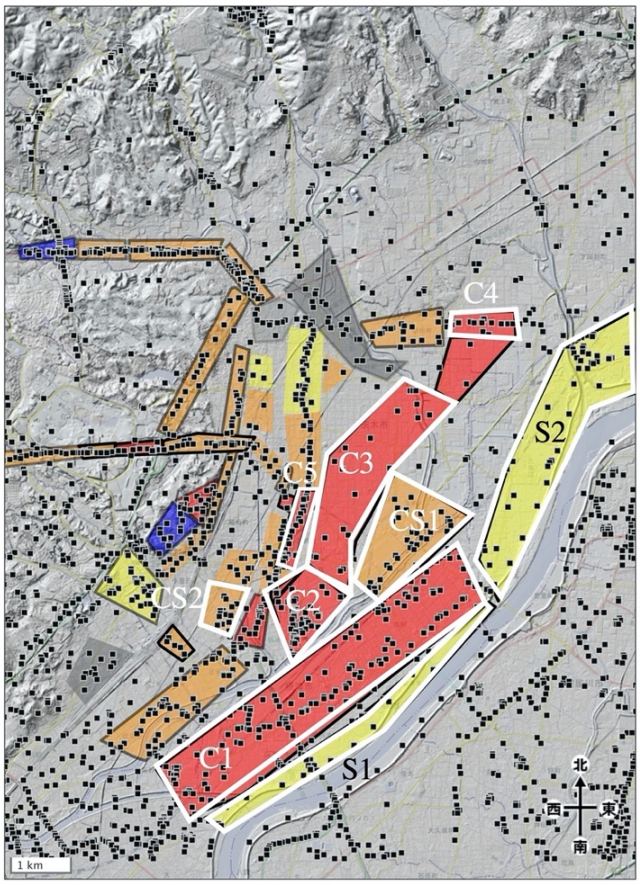



**Figure 5.30** Zoning based on availability of a unique engineering seismic base layer

**Fig. 5.31** explains the meaning of a representative section using an example. **Fig. 5.32** shows the boring data at five locations of the representative section C1. At site 13, site 2136 and site 3244, we can safely assign the gravel or sand layer below 20 m depth as the engineering seismic base layer. However, at site 2527 and site 2542, the gravel or sand layer below 20 m depth is not visible as the boring data is not deep enough. However, if we assign a representative section as shown in **Fig. 5.33**, we can safely assume the entire data at site 2527 and site 2542 as the sediment depth and can use it in the seismic ground response analysis instead of not including the site due to lack of sufficient data.

Thus, in this study, zoning helps in assigning a common engineering seismic base layer for a large area which is helpful in quick modelling of the soil in that zone. However, it also helps include some sites with insufficient data information if they lie in the same zone. In general, a certain fixed layer is defined as the engineering seismic base layer in a case study area, however, that definition may not always reflect the local site condition. An important consideration in this research is to investigate the depth of engineering seismic base layer at each of the boring site. As seen here, except in a few areas where zoning is possible, I investigate and assign an engineering seismic base layer, separately for each site.

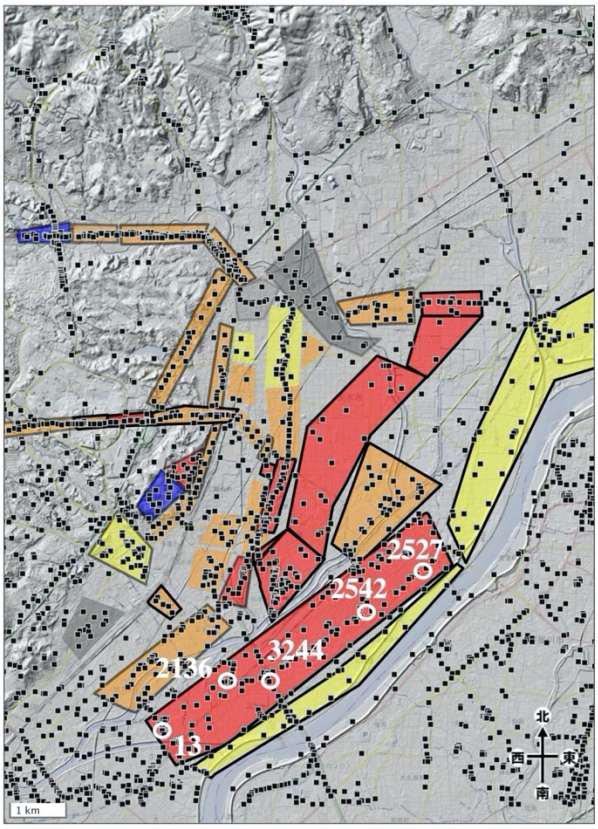



**Figure 5.31** Representative section C1

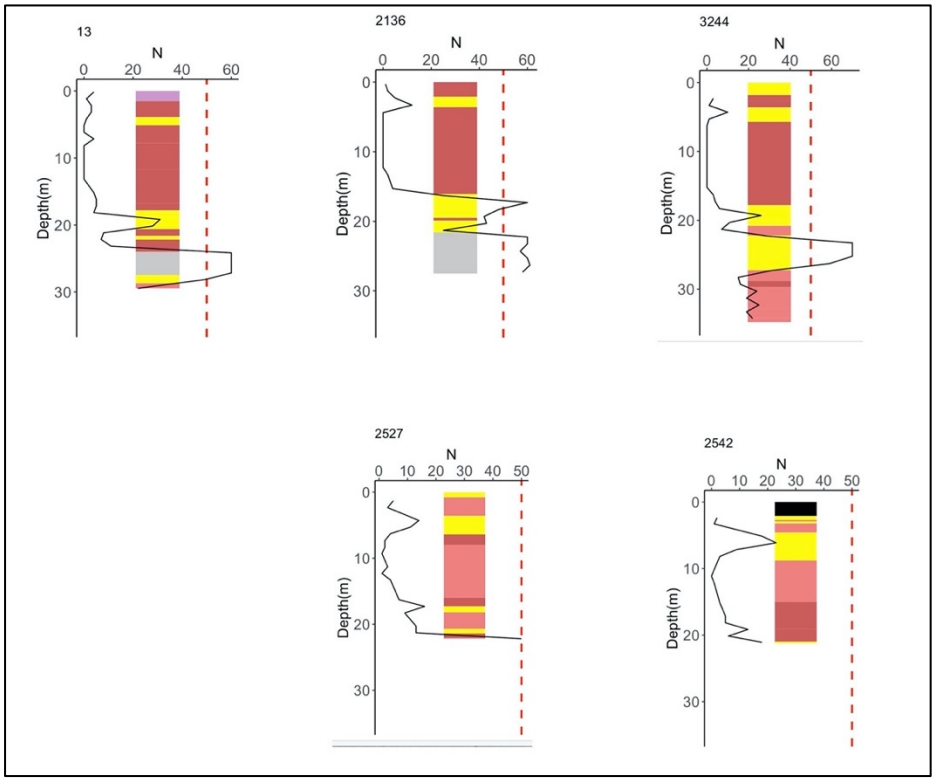



**Figure 5.32** Soil data at multiple locations of representative section C1

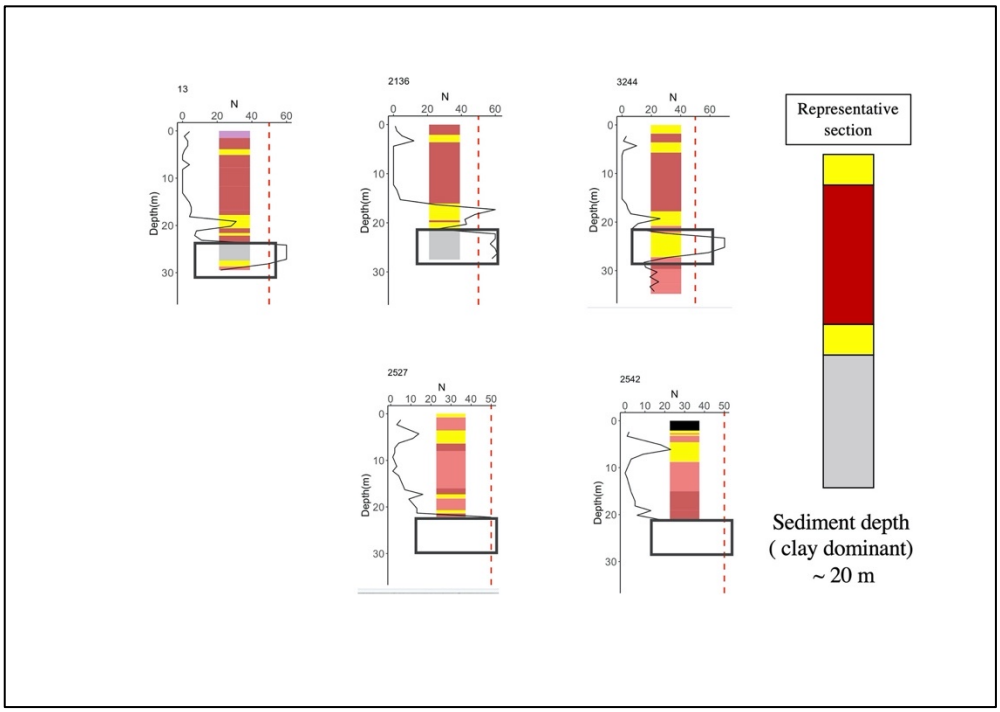



**Figure 5.33** Representative section C1 explained

## 5.5 Calculation of site amplification factor

### 5.5.1 Soil model

The goal of studying the soil structure and assigning an engineering seismic base layer at a site is to model the soil layers at each site. In this study, an earthquake motion is inputted at the engineering seismic base layer and the amplified wave motion at the surface is calculated using seismic ground response analysis. In the J-SHIS map of site amplification factor, the mapped variable is PGV site amplification factor, where PGV is the peak ground velocity. As the objective is to Bayesian update the map resolutions of the J-SHIS map, I also calculate the PGV site amplification factor at every site. I define site amplification factor as the ratio of PGV of input motion at the engineering seismic base layer to the PGV of output motion at the ground surface. **Fig. 5.34** explains the definition of site amplification used in this study.

For the seismic ground response analysis, I employed the multiple reflection theory. The soil material at different layers are assigned an elastic material property and a damping ratio of 2%. However, the shear-wave velocity and density information for all the layers is necessary. The GRI boring data provides only the SPT-*N* value. **Equations (5.3) and (5.4)** are employed to calculate the shear-wave velocity form the SPT-*N* value for different soil types [**18**].

For clay and silt,

$$V_s = 100\, N^{\frac{1}{3}} \tag{5.3}$$

For sand and gravel,

$$V_s = 80\ N^{\frac{1}{3}} \tag{5.4}$$

For sand and gravel if SPT-$N\ \geq 50$, I assumed the Vs as 400m/s.

As for the density, I assign a density of 2 g/cm$^3$ for sand and gravel, and a density of 1.8 g/cm$^3$ for clay and silt. If there is a humus material in the soil, I considered a density of 1.6 g/cm$^3$. **Fig. 5.35** shows the conversion of a SPT-*N* soil data to a layer-wise Vs and density data.

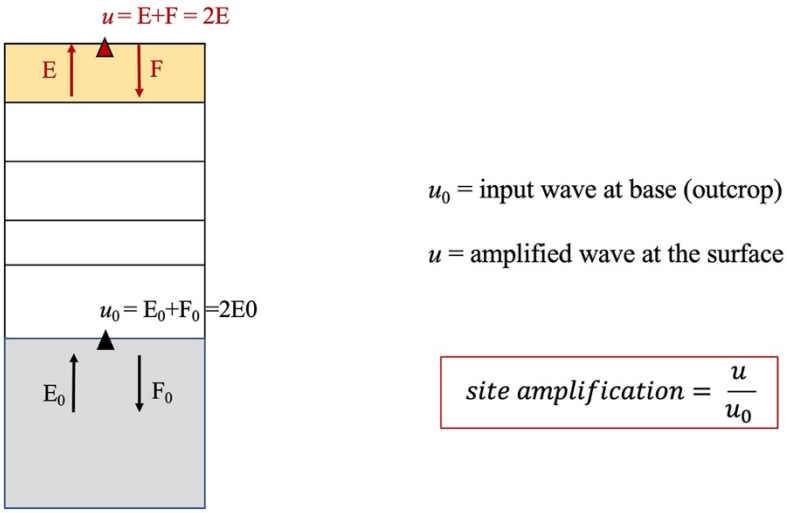



**Figure 5.34** Calculation of site amplification factor using seismic ground response analysis

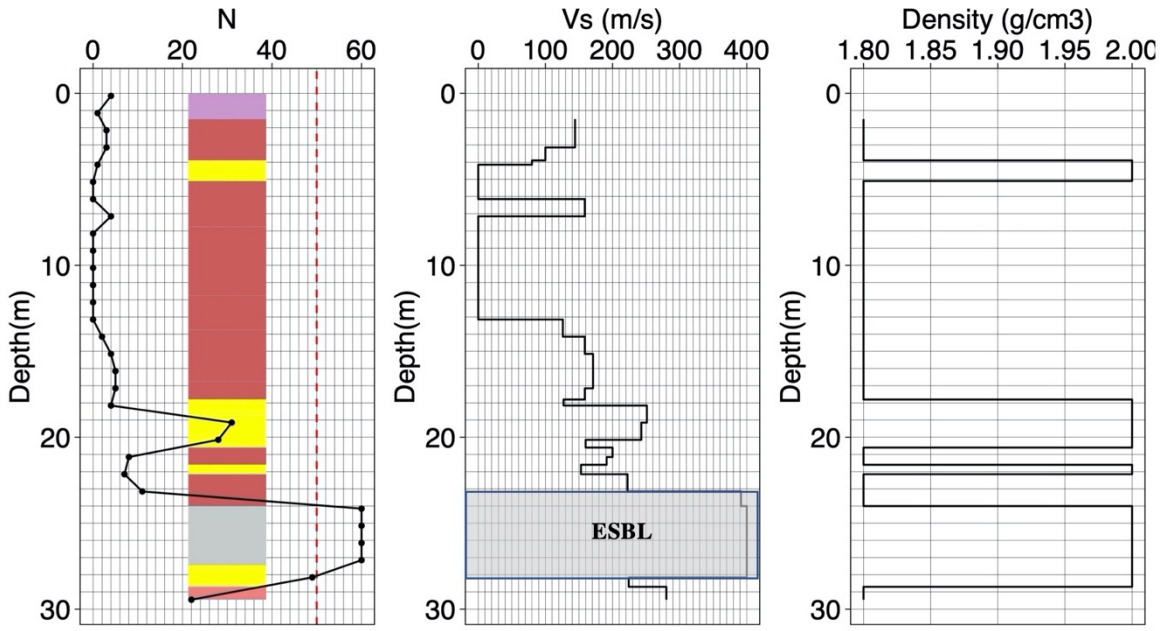



**Figure 5.35** Conversion of SPT-*N* to layer-wise Vs and density data.

ESBL refers to the engineering seismic base layer.

In this study, a sand or gravel layer with shear-wave velocity of 400 m/s is considered as a good candidate for engineering seismic base layer. Although the preferred thickness is around 5 m, however, the actual situation at each site varies based on how the layer continuity is in the neighborhood and how much variation exists in the neighborhood, as discussed in the previous sections.

### 5.5.2 Input ground motion

Outcrop earthquake motion observed at ground surfaces with an S-wave velocity of 400 m/s, can be used as the input earthquake ground motions for our analysis. Amongst the K-NET and KiK-net earthquake data of the National Research Institute for Earth Science and Disaster Resilience (NIED), I selected stations with an average S-wave velocity up to the depth of 5 m (Vs5) in the range of 400~700 m/s. Amongst the earthquake records obtained at these stations, we selected those records with an measured seismic intensity of 4.5 ~ 5.0. **Tables 5.2 and 5.3** lists 21 such earthquake events used in this analysis. **Tables 5.4 and 5.5** lists the PGV and PGA of the input ground motions used in this study.

**Table 5.2** List of selected earthquake events (K-NET)

| Assigned Site Code | NIED Site Code | Vs5 (m/s) | Event Time | Horizontal 2-component PGA ($cm/s^2$) | Measured Seismic Intensity (SI) | Earthquake Type |
|---|---|---|---|---|---|---|
| AKT017 | **AKT017** | 474 | 2008/ 6/14 08:43 | 241 | 4.54 | 内陸地殻内 |
| FKS015A | **FKS015** | 484 | 2011/ 3/11 14:46 | 278 | 4.99 | プレート境界 |
| FKS015B | **FKS015** | 484 | 2011/ 4/11 17:16 | 179 | 4.53 | 内陸地殻内 |
| FKS031 | **FKS031** | 426 | 2011/ 7/31 03:54 | 320 | 4.62 | プレート境界 |
| IWT008 | **IWT008** | 620 | 2011/ 3/11 14:46 | 385 | 4.98 | プレート境界 |
| IWT019A | **IWT019** | 656 | 2003/ 5/26 18:24 | 352 | 4.78 | スラブ内 |
| IWT019B | **IWT019** | 656 | 2011/ 4/ 7 23:32 | 174 | 4.82 | スラブ内 |
| IWT019C | **IWT019** | 656 | 2012/ 3/27 20:00 | 144 | 4.73 | 内陸地殻内？ |
| IWT023 | **IWT023** | 434 | 2008/ 7/24 00:26 | 333 | 4.70 | スラブ内 |
| NAR007 | **NAR007** | 422 | 2016/11/19 11:48 | 404 | 4.52 | スラブ内 |
| SAG001 | **SAG001** | 492 | 2005/ 3/20 10:53 | 330 | 4.51 | 内陸地殻内 |
| SMN015 | **SMN015** | 470 | 2000/10/ 6 13:30 | 268 | 4.86 | 内陸地殻内 |
| TCG002 | **TCG002** | 408 | 2011/ 3/11 14:46 | 154 | 4.57 | プレート境界 |
| YMG014 | **YMG014** | 400 | 2001/ 3/24 15:28 | 142 | 4.72 | スラブ内 |

**Table 5.3** List of selected earthquake events (KiK-net)

| Assigned Site Code | NIED Site Code | Vs5 (m/s) | Event Time | Horizontal 2-component PGA ($cm/s^2$) | Measured Seismic Intensity (SI) | Earthquake Type |
|---|---|---|---|---|---|---|
| HRSH07 | **HRSH07** | 580 | 2014/ 3/14 02:07 | 177 | 4.64 | スラブ内 |
| IWTH09 | **IWTH09** | 440 | 2008/ 7/24 00:26 | 531 | 4.94 | スラブ内 |
| IWTH17 | **IWTH17** | 484 | 2008/ 7/24 00:26 | 402 | 4.53 | スラブ内 |
| IWTH23A | **IWTH23** | 400 | 2008/ 7/24 00:26 | 474 | 4.77 | スラブ内 |
| IWTH23B | **IWTH23** | 400 | 2011/ 4/ 7 23:32 | 542 | 4.98 | スラブ内 |
| IWTH25A | **IWTH25** | 430 | 2008/ 6/14 09:20 | 784 | 4.94 | 内陸地殻内 |
| IWTH25B | **IWTH25** | 430 | 2008/ 6/14 23:42 | 944 | 4.86 | 内陸地殻内 |

**Table 5.4** List I of selected input ground motions. NS and EW refers to the horizontal components of the K-NET earthquake record.

| SL. No. | Assigned Site Code | PGA ($cm/s^2$) | PGV (cm/s) |
|---|---|---|---|
| 1 | AKT017.NS | 222.95 | 12.558 |
| 2 | AKT017.EW | 136.736 | 8.97 |
| 3 | FKS015A.NS | 275.241 | 14.221 |
| 4 | FKS015A.EW | 210.558 | 21.801 |
| 5 | FKS015B.NS | 172.866 | 7.181 |
| 6 | FKS015B.EW | 105.272 | 10.574 |
| 7 | IWT008.NS | 241.035 | 11.453 |
| 8 | IWT008.EW | 323.654 | 16.548 |
| 9 | IWT019A.NS | 256.252 | 7.172 |
| 10 | IWT019A.EW | 285.879 | 13.916 |
| 11 | IWT019B.NS | 145.053 | 10.135 |
| 12 | IWT019B.EW | 145.927 | 10.579 |
| 13 | IWT019C.NS | 101.93 | 6.673 |
| 14 | IWT019C.EW | 134.652 | 11.35 |
| 15 | FKS031.NS | 151.712 | 5.207 |
| 16 | FKS031.EW | 312.83 | 11.265 |
| 17 | IWT023.NS | 324.649 | 6.368 |
| 18 | IWT023.EW | 282.685 | 13.304 |
| 19 | NAR007.NS | 236.785 | 5.22 |
| 20 | NAR007.EW | 347.023 | 7.563 |

**Table 5.5** List II of selected input ground motions. NS and EW refers to the horizontal components of the K-NET earthquake record. NS2 and EW2 refers to the horizontal components of the KiK-net earthquake record.

| SL. No. | Assigned Site Code | PGA ($cm/s^2$) | PGV (cm/s) |
|---|---|---|---|
| 21 | SAG001.NS | 117.668 | 7.766 |
| 22 | SAG001.EW | 329.744 | 10.089 |
| 23 | SMN015.NS | 151.144 | 20.458 |
| 24 | SMN015.EW | 267.493 | 13.604 |
| 25 | TCG002.NS | 104.82 | 7.623 |
| 26 | TCG002.EW | 150.356 | 9.958 |
| 27 | YMG014.NS | 132.496 | 11.342 |
| 28 | YMG014.EW | 129.309 | 13.641 |
| 29 | HRSH07.NS2 | 176.852 | 8.494 |
| 30 | HRSH07.EW2 | 167.486 | 6.827 |
| 31 | IWTH09.NS2 | 421.565 | 9.799 |
| 32 | IWTH09.EW2 | 524.084 | 13.471 |
| 33 | IWTH17.NS2 | 309.531 | 9.157 |
| 34 | IWTH17.EW2 | 324.874 | 10.72 |
| 35 | IWTH23A.NS2 | 399.585 | 7.737 |
| 36 | IWTH23A.EW2 | 399.121 | 11.244 |
| 37 | IWTH23B.NS2 | 507.32 | 10.187 |
| 38 | IWTH23B.EW2 | 483.476 | 11.419 |
| 39 | IWTH25A.NS2 | 781.934 | 21.905 |
| 40 | IWTH25A.EW2 | 210.323 | 5.279 |
| 41 | IWTH25B.NS2 | 344.687 | 6.097 |

### 5.5.3 Site-specific variation of amplification factor

One site doesn't have a unique PGV site amplification factor. As the input motion changes, the amplitude and phase of the wave changes, affecting how it interacts with the soil layers and ultimately affecting the amplified output wave. This change in PGV site amplification factor with change in input motion gives rise to a site-specific variation. However, this variation will change from site to site, giving rise to a spatial variation of the PGV site amplification factor. In UPM, both these variations of data will be incorporated. The conventional maps fail to incorporate such an information of local variation in the mapping process. **Table 5.6** shows how the PGV site amplification factor varies with the input motion characteristics at the example site introduced in **Fig 5.35.**

**Table 5.6** Site-specific variation of PGV site amplification factor

| Input Motion at Base | PGV(cm/s) ($2E_0$) | Output PGV(cm/s) at surface (E+F) | PGV Amplification (E+F/ $2E_0$) |
|---|---|---|---|
| A | 12.55 | 17.24 | 1.37 |
| B | 8.97 | 16.16 | 1.80 |
| C | 14.22 | 23.49 | 1.65 |
| D | 21.8 | 27.63 | 1.26 |
| E | 11.45 | 26.85 | 2.34 |
| F | 16.55 | 31.91 | 1.92 |

## 5.6 Observed data for UPM

In this study, I use the PGV site amplification factors calculated for the 41 input earthquake motions at each of the 485 boring sites in the case study area as the observation data for UPM. **Figs 5.36 (a) and (b)** shows the mean and standard deviation of the observation data. The color scale for the mean is maintained the same as the J-SHIS map of site amplification factor. However, as the J-SHIS map of site amplification factor is based on 250 m×250 m meshes, the 485 boring sites will need to be assigned to their appropriate meshes. **Fig. 5.37** shows the assignment of the boring sites to the appropriate meshes. **Figs 5.38 (a) and (b)** show the mesh-mean and mesh-standard deviation of observation data at each of the 250 m×250 m mesh in the case study area.

In chapters 3 and 4, I worked with point data for the UPM mapping. However, in this chapter, I will prepare mesh based UPM maps. **Fig. 5.39** shows the initial neighborhood definition for the UPM. As neighbors, I defined those meshes which share a border with the mesh under consideration. However, the neighborhood changes as the *c*-value is changed and the best model is selected based on model evaluation.

In the next section, the posterior probabilities (Bayesian updated results) estimated using the informative prior (J-SHIS map of site amplification factor) and observation data (PGV site amplification factors calculated from the soil boring data) are discussed.

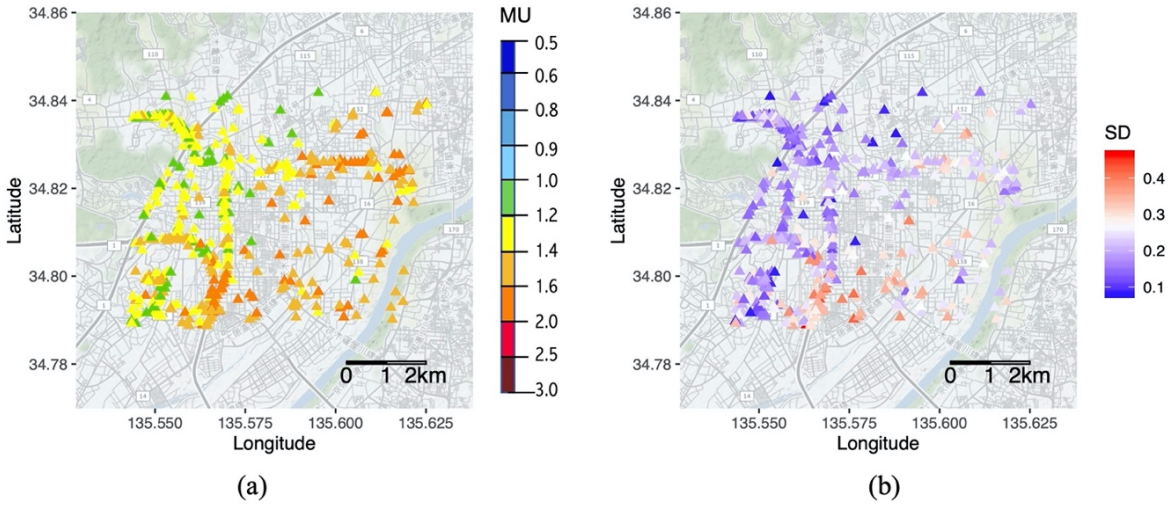



**Figure 5.36** Distribution of average (MU) and standard deviation (SD) of observation data for UPM

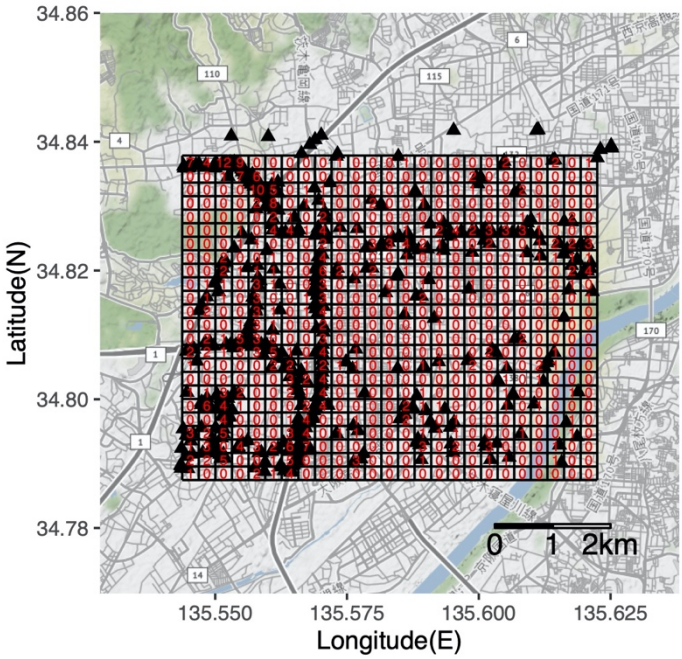



**Figure 5.37** Assignment of boring sites to the meshes of size 250 m×250 m

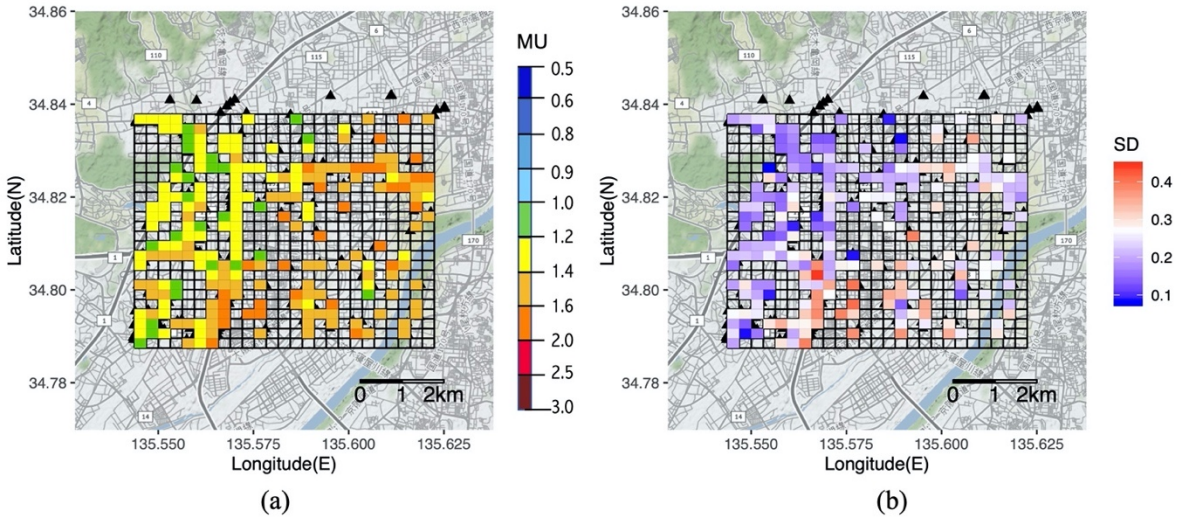



**Figure 5.38** Mesh-wise distribution of average (MU) and standard deviation (SD) of observation data for UPM

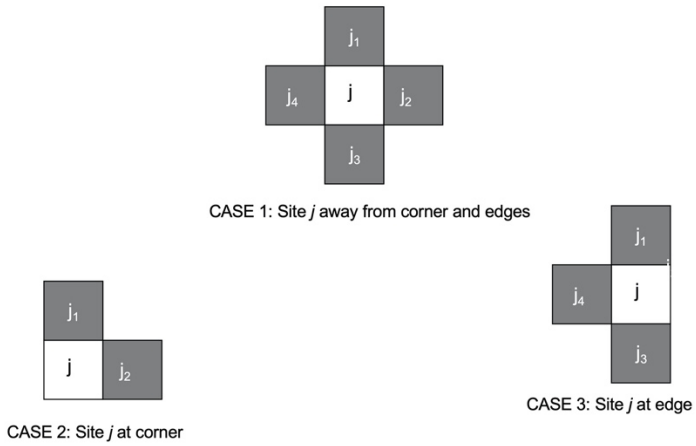



**Figure 5.39** Initial neighborhood definition for mesh based UPM

## 5.7 Result: Updated J-SHIS map and comparison with its original counterpart

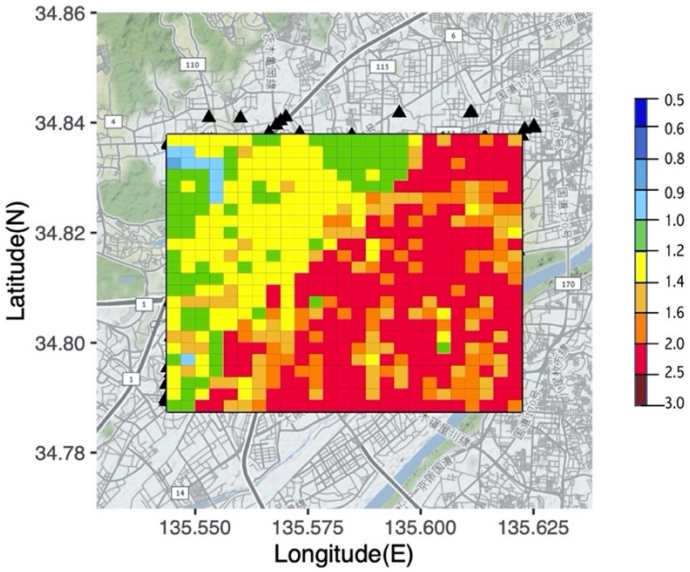



**Figure 5.40** Updated J-SHIS map of site amplification factor

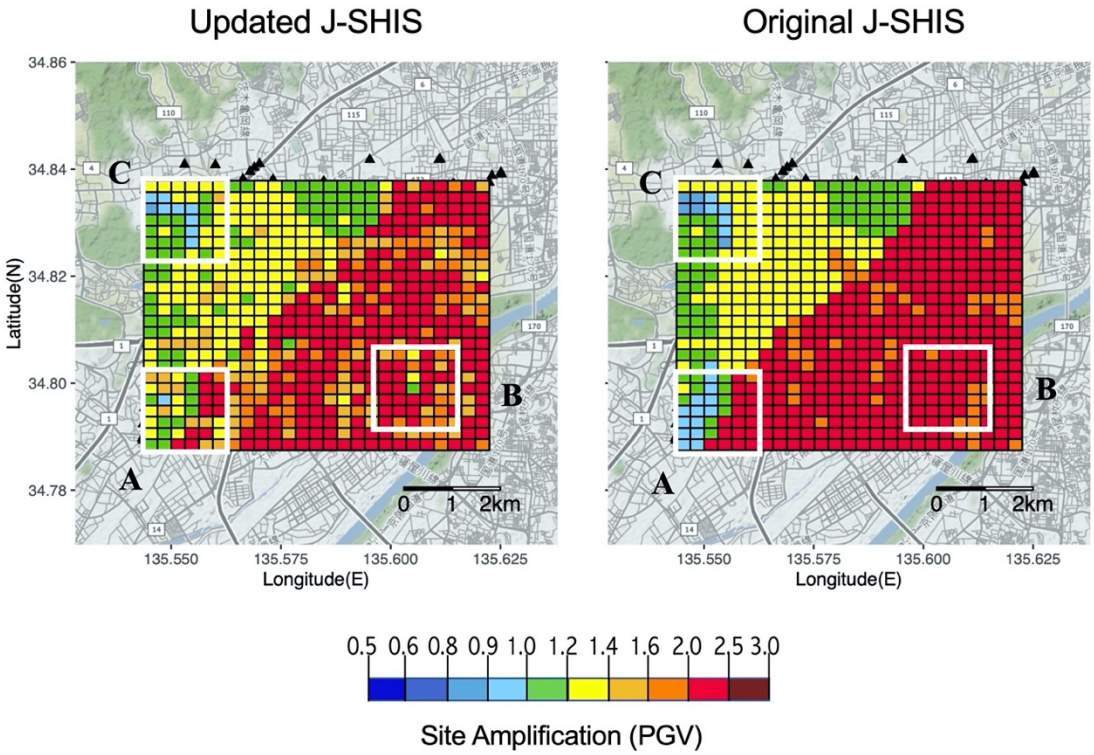



**Figure 5.41** Comparison of updated and original J-SHIS map of site amplification factor

**Fig. 5.40** shows the Bayesian updated J-SHIS map of site amplification factor. In order to understand the significance of the updated map resolutions, I compare it with the original J-SHIS map resolutions as shown in **Fig. 5.41**. When both the maps are compared, the first thing that can be observed is that for meshes where local boring data is available, the map resolutions are updated and reflect the local information. However, for meshes where local boring data is not available, the original J-SHIS map resolutions are maintained because of the strong prior information.

In order to understand the significance of the updated map resolutions more clearly, I draw attention to three zones marked by A, B and C in Fig. 5.41.

In Zone A, the J-SHIS map shows a high amplification area (in red) and low amplification area (in blue) situated right next to each other. However, in the updated J-SHIS map the contrast is significantly reduced and a smoother transition of map resolution colors is shown. In Zone B, the J-SHIS map shows almost no contrast in the map resolutions. However, the updated J-SHIS map resolutions show the presence of a significant low amplification area (in green). In Zone C, both J-SHIS and its updated version, show a smooth transition of map resolution colors.

In order to explain these similarities and dissimilarities discussed above, I investigate the nature of boring data in the above-mentioned zones as shown in **Fig. 5.42**.

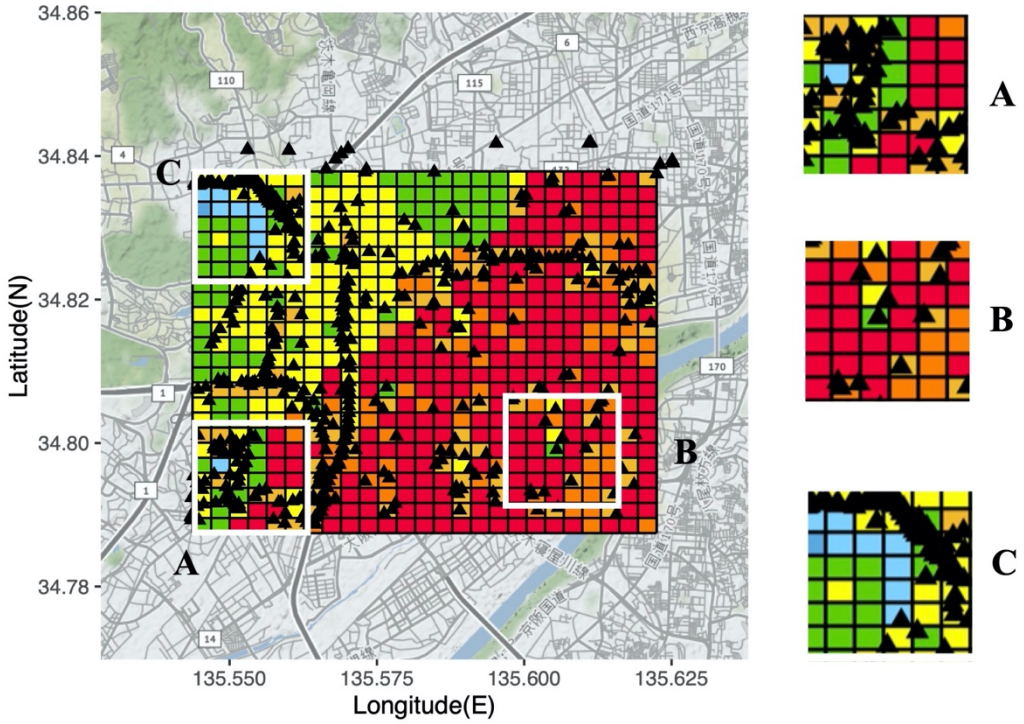



**Figure 5.42** Location of boring data (▲) in zones A, B and C of updated J-SHIS map.

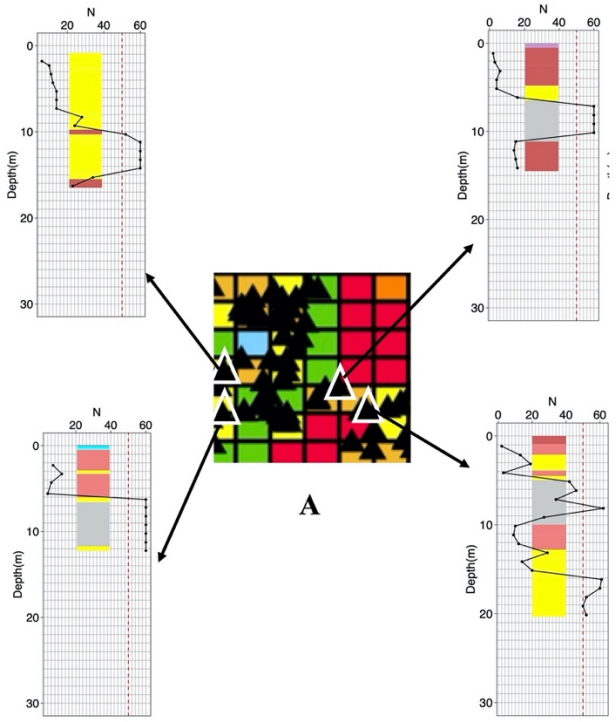



**Figure 5.43** Detailed boring data of zone A in case study area

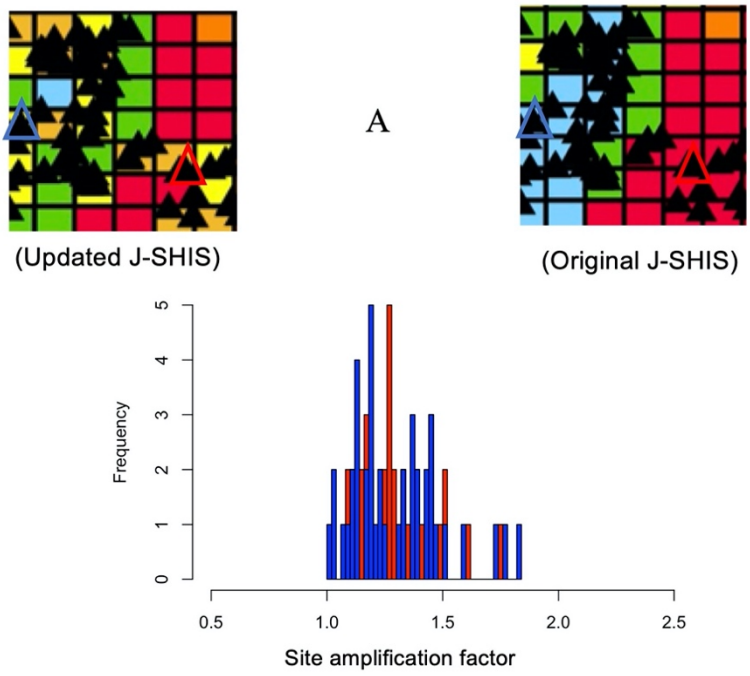



**Figure 5.44** Histogram of site amplification factors for two sites (red and blue) in zone A

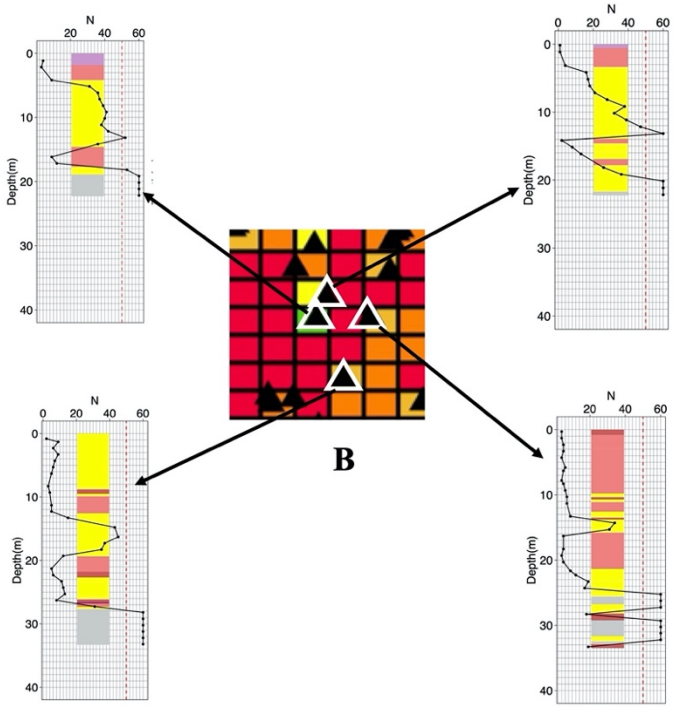



**Figure 5.45** Detailed boring data of zone B in case study area

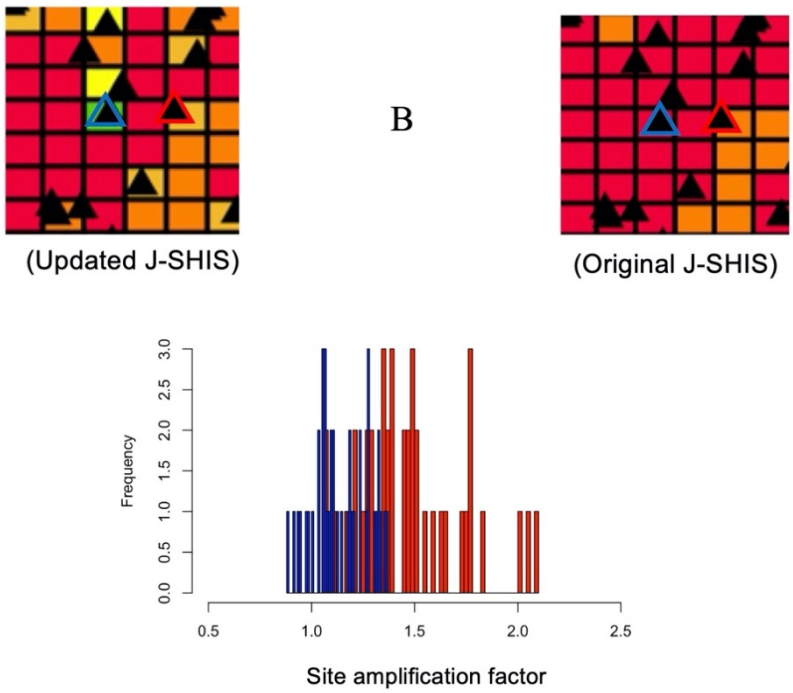



**Figure 5.46** Histogram of site amplification factors for two sites (red and blue) in zone B

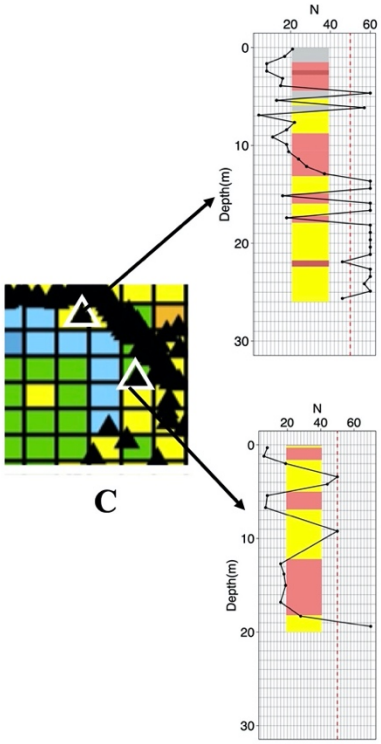



**Figure 5.47** Detailed boring data of zone C in case study area

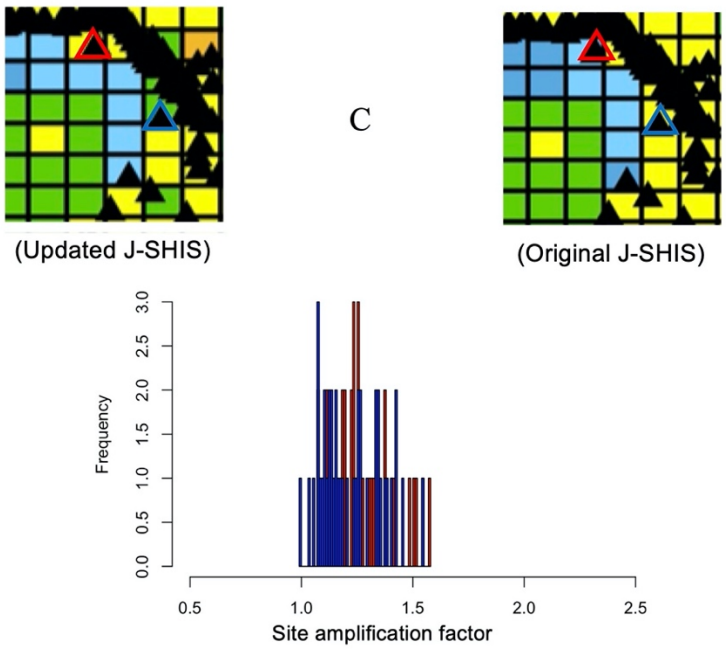



**Figure 5.48** Histogram of site amplification factors for two sites (red and blue) in zone C

I first investigate zone A in **Fig 5.43**. Four boring data (two in original J-SHIS high amplification area and two in original J-SHIS low amplification area) are shown in detail. It is observed that there isn't a significant change in the sediment thickness in the two areas. The histogram of site amplification factors for two sites in low and high amplification area in original J-SHIS map shows high overlapping (**Fig. 5.44**). Thus, the high spatial resolution in the original J-SHIS map cannot be explained statistically. And hence the updated map resolutions with lower contrast is more reliable based on the available data.

On investigating boring data of zone B in **Fig 5.45**, two sites with substantially low sediment thickness are visible. Also, this is a zone of transition from the sand dominated layer to the clay dominated layer as indicated by our zoning (**Fig. 5.29**). The updated J-SHIS map resolutions, thus, highlights the significant difference of these sites with respect to their neighborhood unlike the original J-SHIS map resolutions which considered the whole area as uniform. The histogram of site amplification factors for the two neighboring sites also show very low overlapping, and hence validates the spatial resolution of the updated J-SHIS map on the basis of data uncertainty (**Fig. 5.46**).

In Zone C, the spatial resolution of the updated J-SHIS map is low as the number of data points is significantly low in the area (**Fig. 5.47**). The data is not enough to highlight significant differences in the area and thus it has a high predictive error. The histogram of site amplification factors for the two sites also show very high overlapping (**Fig. 5.48**).

Thus, the updated J-SHIS map reflects the data information and highlights significant differences when such a difference exists and for cases with low data where information cannot be extracted, a low spatial resolution (smooth mapping) is introduced adding more reliability (in statistical terms) in comparison to original J-SHIS map.

## 5.8 Discussion

The reliability issue with J-SHIS map resolutions at the local scale is widely acknowledged. In this study, I addressed this reliability issue by incorporating local soil boring data into the mapping process. The map resolutions at the meshes with data were updated. However, more data will be necessary to update each and every mesh of the J-SHIS map.

During the 2018 Northern Osaka earthquake, in Ibaraki city of Osaka, it has been observed that certain areas with low J-SHIS site amplification factor suffered more damage whereas certain areas with high J-SHIS site amplification factor observed little or no damage at all [**19,20**]. Damage was observed in our case study area too. A comparison of the earthquake damage distribution in the case study area with the updated J-SHIS map resolutions was attempted in **Fig. 5.49**.

**Equation (5.6)** defines the damage ratio as follows [20],

$$Damage\ ratio\ (in\ \%) = \frac{The\ number\ of\ residence\ damage\ certificates}{The\ \ number\ of\ building} \times 100 \qquad (5.5)$$

where the residence damage certificate is proof of extent of the damage of residence. In Japan, after a disaster, people who have their houses damaged by the earthquake will require several certificates of proof in order to be eligible to receive aid.

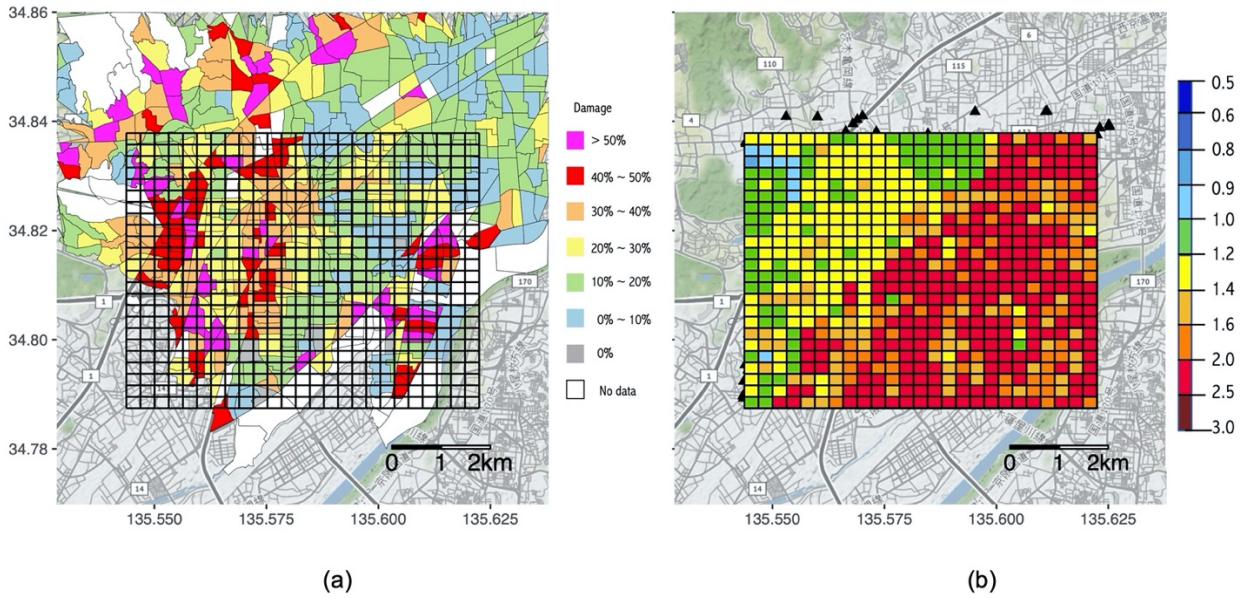



**Figure 5.49** (a) Damage distribution during the 2018 Northern Osaka Earthquake [**20**]
(b) Updated J-SHIS map

There are many short comings of this comparison. In this study, a linear analysis of seismic ground response is performed, and the input earthquake motions are not derived from nearby faults. Damage is not only due to site amplification effect but important factors like source effect, age of buildings, etc. significantly contribute to the damage distribution, which are not considered in this study. Thus, although, no significant discussion on the distribution can be done, but because the map resolutions are now enriched with local soil information, the updated J-SHIS map might be used as a supporting evidence in explaining the damage distribution in the future. An important point to note is that the discrepancies between the observed damage distribution and the J-SHIS map resolutions highlight the need to include more information in the mapping process and our approach of incorporating local soil data could play a significant role in adding more reliability to the spatial resolutions.

## 5.9 Conclusion

The J-SHIS map of site amplification factor is based on AVs30 values broadly assigned to engineering geomorphic units. There is a need to incorporate local (site-specific) information in order to address the reliability issues with the resolutions at local scales.

In this chapter, I apply the framework of UPM to incorporate uncertainties from local data sources and update the map resolutions of J-SHIS map. As a case study area, I chose Ibaraki-Takatsuki area of Osaka. Using seismic ground response analysis, I calculated the site amplification factor from the available boring data in individual meshes. The updated J-SHIS map reflects the data information and highlights significant differences, when such a difference exists, and for situations with low data where information cannot be extracted, a low spatial resolution (smooth mapping) is introduced adding more reliability (in statistical terms) in comparison to original J-SHIS map.

## References


1. J-SHIS Maps: http://www.j-shis.bosai.go.jp/map/?lang=en (accessed on 2020.10.07)
2. J-SHIS Maps user manual: http://www.j-shis.bosai.go.jp/map/JSHIS2/man/en/ (accessed on 2020.10.29)
3. J-SHIS: What is the 250m-mesh code? http://www.j-shis.bosai.go.jp/en/faq-250mmesh#more-42 (accessed on 2020.10.29)
4. Wakamatsu, K., & Matsuoka, M. (2013). Nationwide 7.5-arc-second Japan engineering geomorphologic classification map and Vs30 zoning. Journal of Disaster Research, 8(5), 904-911.
5. Matsuoka, M. and Wakamatsu, K. (2008): "Site Amplification Capability Map based on the 7.5-arc-second Japan Engineering Geomorphologic Classification Map", National Institute of Advanced Industrial Science and Technology, Intellectual property management, No.H20PRO-936.
6. Fujimoto, K., & Midorikawa, S. (2006). Relationship between average shear-wave velocity and site amplification inferred from strong motion records at nearby station pairs. Journal of Japan Association for Earthquake Engineering, 6(1), 11-22.
7. Yoshida, N. (2015). Seismic ground response analysis. Dordrecht: Springer Netherlands.
8. Seed, H. B. & Idriss, I. M. (1970). Soil moduli and damping factors for dynamic response analyses, report no. EERC70-10, Earthquake Engineering Research Center, *University of California, Berkeley*, 40p
9. Hardin, B. O., & Drnevich, V. P. (1972). Shear modulus and damping in soils: design equations and curves. *Journal of Soil Mechanics & Foundations Div*, 98(sm7).
10. Hardin, B. O., & Drnevich, V. P. (1972). Shear modulus and damping in soils: measurement and parameter effects. *Journal of Soil Mechanics & Foundations Div*, 98(sm6).
11. Schnabel, P. B. (1972). SHAKE: A computer program for earthquake response analysis of horizontally layered sites. *EERC Report 72-12, University of California, Berkeley.*

12. Toki, K. (1981). Seismic response analysis of structures. New series of civil engineering, vol 11, Gihodo Shuppan, 250 pp (in Japanese)
13. Japan Society for Natural Disaster Science (2002) Subject-book on disaster prevention, Tsukiji Shobo, 543 pp
14. Amachi, F. (2009). Concept of bedrock and ground characteristics, Lecture on fundamentals of strong motion geotechnology, Seismological Society of Japan
15. Port and Harbors Bureau, Ministry of Transport (eds) (1999) Technical standards and commentary of port and harbor facilities in Japan. The overseas coastal area development institute of Japan (in Japanese)
16. Japan Road Association. (2002) Specifications for highway bridges. Part V, Seismic design, Maruzen, 317 pp
17. Construction Ministry (2000) On structural calculation standard for structural safety of high-rise building, ministerial announcements 1461 (in Japanese)
18. Japan Road Association. (1980). Specifications for highway bridges, Part V, Seismic design
19. Goto, H. (2018). Ground motion characteristics during the 2018 northern Osaka earthquake, The 31st KKHTCNN Symposium on Civil Engineering, Nov. 22-24.
20. 浅野晃太，奥村与志弘，澤田純男，後藤浩之 (2020).2018 年大阪府北部の地震における高槻市・茨木市の建物被害分布と要因に関する地理学的考察，第 40 回地震工学研究発表会，Oct. 1-2.

# CHAPTER 6

# CONCLUSION

In this research, I address the important issue of incorporating uncertainty information in the map resolutions for better decision-making. In literature, I hardly came across any research addressing this issue of incorporating data uncertainty in the map resolutions. The topic is important as the statistical significance of this difference in neighboring values is directly ungraspable without any information on the data uncertainty. The inability of conventional maps to statistically signify the difference in mapped values, raises a question on its use for reliable decision-making process.

In the first study, a relation between the site-specific uncertainty and spatial uncertainty in the framework of a hierarchical Bayesian model is introduced. The idea is to make the map resolutions uniform or smooth at zones of high data uncertainty. I named this mapping methodology as Uncertainty Projected Mapping (UPM). The proposed UPM methodology was validated with both numerical experiments and real data from a very dense seismic array. The UPM results were found to reflect the site-specific uncertainties in the map resolutions. The detailed visual (mapping) on significantly different observations expected between the sites with low standard deviation and rough visual (mapping) on insignificant observations between the sites, was enhanced in the UPM maps. The UPM methodology could spatially interpolate and estimate values at the missing sites. The values at the missing sites are naturally estimated based on the spatial structure introduced by the CAR model. The mapping results were compared with a conventional mapping technique called Kriging. It was observed that unlike the UPM values which are sensitive to the variation of data uncertainty, the Kriging values are not affected by the change in data uncertainties.

In the second study, I investigate the information-dependent characteristics of UPM. It is observed that as more and more information become available, UPM starts approaching the conventional mapping. This characteristic hints at the strength of UPM when less information is available. I investigate this characteristic in detail and utilize it to propose a parameter to measure the change in map resolutions with increasing information, which was applied for quantification of data saturation in mapping spatial data. The results show that the optimum number of data which is deemed enough to extract useful information depends on available dataset. This study also establishes the strength of UPM when less information is available.

In the third study, I use the framework of UPM to incorporate local uncertainty information and update the map resolutions of J-SHIS map of site amplification factor in Ibaraki-Takatsuki area of Osaka. The J-SHIS map of site amplification factor is based on AVs30 values broadly assigned to engineering geomorphic units. There is a need to incorporate local (site-specific) information in order to address the reliability issues with the resolutions at local scales. Using seismic ground response analysis, I calculated site amplification factor from the available boring data in individual meshes. Further, using the Bayesian framework of the UPM, I updated the existing map resolutions of J-SHIS map. The updated J-SHIS map reflects the data information and highlights significant differences, when such a difference exists, and for situations with low data where information cannot be extracted, a low spatial resolution (smooth mapping) is introduced adding more reliability (in statistical terms) in comparison to original J-SHIS map.

The results in all the three studies primarily establish that that map resolutions determined by data uncertainty has a statistical significance and are a better representation of the whole data at a site.

## ACKNOWLEDGEMENTS

Working on this research problem which deals with uncertainty has been an interesting and extremely rewarding journey. I would like to thank several people who contributed to the final outcome in many different ways. First and foremost, I am deeply indebted to my supervisors, Prof. Sumio Sawada and Associate Prof. Hiroyuki Goto, who have supported me throughout the research with their patience and knowledge whilst allowing me the room to work on my own way. Sawada Sensei's inquisitive nature has always reminded me the fact that there is no end to learning. Goto Sensei have always been more than only a guide and his door was always open for me (literally). I am grateful to both of them for providing me many opportunities to present my work on different national and international platforms.

I owe my special thanks to Kubo Sensei from Hokkaido University, whose book "データ解析のための統計モデリング入門" helped me apply the conditional auto regression models practically. It was very crucial for my research and without his book it would have been very difficult. I am also grateful to Prof. Junji Kiyono, co-supervisor of my Doctoral course, for his important comments which made me see my research results from an angle which was new for me.

In my daily laboratory life, I have been blessed with a very friendly group of colleagues. In the last three years, I spent a lot of time with Daiki Yamashita and Kodai Kato, both inside and outside the lab. It has always been a refreshing experience spending time with them. I always enjoyed the deep conversations with Jun Kurima. I am grateful to Yomi Harada for letting me join her centrifuge experiments which helped me gain some interest in a new research area. I am thankful to Kota Asano for sharing his research results on the 2018 Northern Osaka earthquake and clarifying many of my queries. Our lab secretary, Mrs. Miho Mori who has gone beyond her way many times to help me get official documents for foreign travel at very short notices. I am especially thankful to Prof. Sumio Sawada without whose words of motivation I wouldn't have had the zeal to continue my learning Japanese and today, thanks to his advice, I can manage to communicate in Japanese.

I am very grateful to my friends in Japan and back in India without whom it would have been difficult to survive away from home. I would like to extend a heartfelt thank you to Sachiko Takahashi, Elif Berna Var, Ashif Equbal, Soumya Sethi, Kabya Kaushik, Tala Bakhtiary, Viliame Waqalevu, Sainimere Veitata, Chika Kondo, Shreshth Sapra, Kanika Bimrah and Jasleen Kaur for being an integral part of my emotional support system in Japan. Thank you Swanak Datta, my dear friend in India, for being always available to call and discuss issues, both personal and professional.

And, finally, I thank my parents for supporting me throughout all these years of study and life in general. I dedicate the doctoral thesis to them.