A geo-statistical framework to reduce uncertainty in predictions of V_{S30} and other geotechnical variables

C. E. L. Gilder
University of Bristol, Bristol, UK

R. De Risi, F. De Luca, P. J. Vardanega
University of Bristol, Bristol, UK

R. M. Pokhrel
Earth Investigation and Solution Nepal Pvt Ltd, Kathmandu, Nepal

ABSTRACT: Geo-statistical modelling challenges arise when regional models in data-scarce regions are required. This is relevant when evaluating and dealing with geotechnical uncertainty in earthquake engineering applications. From a geotechnical earthquake engineering perspective, V_{S30} (shear-wave velocity in the upper 30m of soil) predictions must cover regional scale study areas. A systematic lack of data can significantly limit obtaining a satisfactory dataset for accurate soil amplification definition. This paper implements a novel geo-statistical framework to create V_{S30} mapping. Such a framework employs an approach of Bayesian Kriging, implemented initially within the context of petroleum reservoir modelling and recently applied to the case of Kathmandu Valley. The approach uses primary and secondary data to apply either debiasing or declustering to deal with typical issues of data availability in data-scarce regions. The multidisciplinary use of this analysis method provides an assessment of uncertainty, and an informed quantification of geotechnical parameters where traditional statistical methods may not produce sufficiently acceptable results. In this paper, a new set of secondary data using the case study of Kathmandu Valley (Nepal) is employed to demonstrate the flexibility of the geo-statistical framework developed in a previous study by the same authors.

1. INTRODUCTION

Using geo-statistics is a standard methodology for predicting parameters at regional level where measurements are scarce. As Kriging is a linear estimator, it is constrained by the data range of the original measurements (Deutsch & Journel, 1998). Usually, this would be a good assumption (capturing a dataset variability about the mean), yet, when a limited range of measurements is used for a large-scale prediction, Kriging, in the traditional sense, may not capture the full range of possible scenarios. Estimation can become even more challenging when data scarcity leads to sampling bias.

This is the context of the geo-statistical framework developed by Gilder *et al.* (2022) where the combined use of secondary data and Bayesian Kriging enables an improved estimation. This framework was originally developed for the estimation of the shear-wave velocity in the upper 30m of depth (V_{S30}) (see also the thesis of Gilder (2022) for more details on the development of the framework). In this study, further use of the framework is presented for additional geotechnical parameters, using the same case study of the Kathmandu Valley in Nepal.

2. V_{S30} PREDICTION

V_{S30} can be estimated from correlation with topographic gradient (slope) obtained from Digital Elevation Models (DEM) (Wald & Allen 2007; Allen & Wald 2009). This method is used by the United States Geological Survey (USGS) to develop shake maps (Wald *et al.* 2006). From an engineering geology perspective, this relationship simulates the effect expected from soils with different friction angles. The basis of the used predictors is that soils on steep slopes must indicate rock with corresponding high V_{S30}. Where V_{S30} is not measured, Eurocode 8 (CEN 2004) suggests using the parameter SPT-N. Other proxies for V_{S30} include the geology-based model (Wills & Clahan 2006) or terrain-based models (Yong *et al.* 2012). A hybrid combination of these methodologies is commonly used to include the relative benefits of the different methods (e.g., Thompson *et al.* 2014; Stewart *et al.* 2014; Wills *et al.* 2015).

Uncertainty may be associated with measurements, and the possibility of sampling bias (preferentially testing particular sediments, or near seismic stations or for other reasons caused by accessibility issues) has led to the hypothesis that distributions of V_{S30} built from regional databases may potentially be skewed (Mital *et al.* 2021).

The geo-statistical framework developed by Gilder *et al.* (2022) to predict V_{S30} uses both primary and secondary data and implements the multi-gaussian Bayesian updating technique presented by Deutsch & Zanon (2007) building upon the robust kriging approach developed in De Risi *et al.* (2021). When using the methodology, the term *primary* is given to the variable being predicted in the Kriging (which is scarce geospatially). *Secondary* is a densely sampled variable that can provide information for predicting the primary variable. This differs from previous V_{S30} geospatial estimation efforts as the method can fully capture both local scale measurements and regional scale trends.

The case study used in this research was made possible by the SAFER geodatabase for the Kathmandu Valley (Gilder *et al.* 2020), assembled as part of the Global Challenges Research Fund project Seismic Safety and Resilience of Schools in Nepal (SAFER) (https://www.safernepal.net/) funded by the Engineering and Physical Sciences Research Council. This provided a suitable example of geotechnical uncertainties affecting an earthquake-prone region and seismic hazard estimation. The first section of this paper presents the selection of secondary predictors. The second section uses the geo-statistical framework developed in Gilder *et al.* (2022) to present a Bayesian SPT-N map for the region.

3. METHODOLOGY

The methodology of Gilder *et al.* (2022) is organised into six steps:

1. Selection of primary and secondary variables.
2. Selection of Option A or Option B (data debiasing or declustering).
3. Variable transformation.
4. Simple Kriging of the primary variable.
5. Definition of likelihood distribution using the secondary variable/s.
6. Bayesian updating of Step 4.

Taken from the original paper, Step 6 relies on the Bayesian Updating Equations presented by Deutsch & Zanon (2007):

$$\bar{y}_U = \frac{\bar{y}_L\sigma_P^2 + \bar{y}_P\sigma_L^2}{\sigma_P^2 - \sigma_P^2\sigma_L^2 + \sigma_L^2} \qquad (1)$$

$$\sigma_U^2 = \frac{\sigma_P^2\ \sigma_L^2}{\sigma_P^2 - \sigma_P^2\sigma_L^2 + \sigma_L^2} \qquad (2)$$

where subscript P is for the prior distribution, U is for the updated distribution, and L is for the likelihood. The $\bar{y}_U$ is the updated Kriging output, and $\sigma^2{}_U$ is the accompanying estimation variance for that prediction. The $\bar{y}_P$ is the Simple Kriging result of Step 4 (*i.e.*, mean of the distribution at each location, $\mathbf{u}$), σ^2_P is the accompanying estimation variance of the Simple Kriging prediction, $\bar{y}_L$ is the mean likelihood, and σ^2_L is the likelihood variance built from the secondary variables.

Step 4 involves the development of a variogram. The variogram is created in geostatistics to understand whether samples close together have similar properties. This aspect is left to the modeller's discretion, including the choice of variogram type, *i.e.,* spherical or exponential (*e.g.,* Wackernagel 2010). When using the framework presented here, the variogram must be decided prior to any declustering or debiasing of the primary data. Before using Equations 1 and 2, the original data values of the primary measurements and secondary predictor are transformed to normal space.

In step 2, as explained in Gilder *et al.* (2022), sampling bias in a dataset can be common, as often, geotechnical data are taken at specific points with the intent of answering a particular question. Debiasing of data can be accomplished by incorporating a data trend between primary and secondary variables. A judgement-based bivariate trend is needed to translate the conditional distribution of the primary variable, given a secondary variable, to portions of the distribution where data is unknown. Step 2 (Option A) uses:

$$f_Y(y) = \int_x f_{Y|X}(y|x)\, f_X(x)\, dx \qquad (3)$$

where $f_Y(y)$ is the debiased distribution of the primary variable, $f_{Y|X}(y|x)$ is a conditional distribution between secondary (x) and primary (y) variables, where they are collocated (Deutsch *et al.* 1999). *Collocated* is used to describe where both parameters are available at a particular location, $\mathbf{u}$. $f_X(x)$ is the distribution provided by the secondary data, densely sampled and unbiased, so it provides a better understanding of the entire possible range of y.

Declustering is selected as a debiasing option when a primary dataset is favouring a particular interval of values. This might occur when drilling in a concentrated area exhibiting particularly high or low primary parameter values. Step 2 (Option B) uses declustering in the configuration:

$$y_{d,i} = \alpha_i y_i = \frac{AV_i}{A} y_i \qquad (4)$$

where a weight is assigned to a primary data point (y_i), using the ratio (α_i) between the area of the Voronoi polygon around each data (AV_i) and the total area of the window of interest (A).

4. ANALYSIS

4.1. Secondary predictors

The process of selecting a good secondary predictor is dependent on the strength of correlation with the primary variable. Where V_{S30} is the variable to be predicted, several options for secondary data may be available based on understanding how it may vary. The database developed for the Kathmandu region by the authors (Gilder *et al.* 2020) is used as a source of additional data. Where in a data-scarce region, the option of having a densely populated secondary variable is uncommon, the first step may be to undertake simple Kriging to understand if the resulting distribution of a secondary dataset characterises the region effectively.

Soil grain size was selected as an option due to the understood correlation with V_{S30} values, (i.e., splitting correlation models for grain size is a common approach) (Wair *et al*. 2012). In Figure 1, the updated Kriging estimate from Gilder *et al.* (2022) where slope based V_{S30} estimation and bedrock depth was used as secondary data (Figure 1a) is compared with the percentage of soil passing 0.075 mm sieve (*% fines*) (Figure 1b).

The data points for grain size informing the Simple Kriging are averaged at each location for the entire soil column (ranging between 10 m and 35 m depth). This has resulted in 63 data points (circle points in Figure 1b). However, many points occur at concurrent sites and are not distinguishable or suitable for developing the model at the regional scale. Considering the above limitations and given the distribution in Figure 1b, grain size does not form a good secondary predictor.

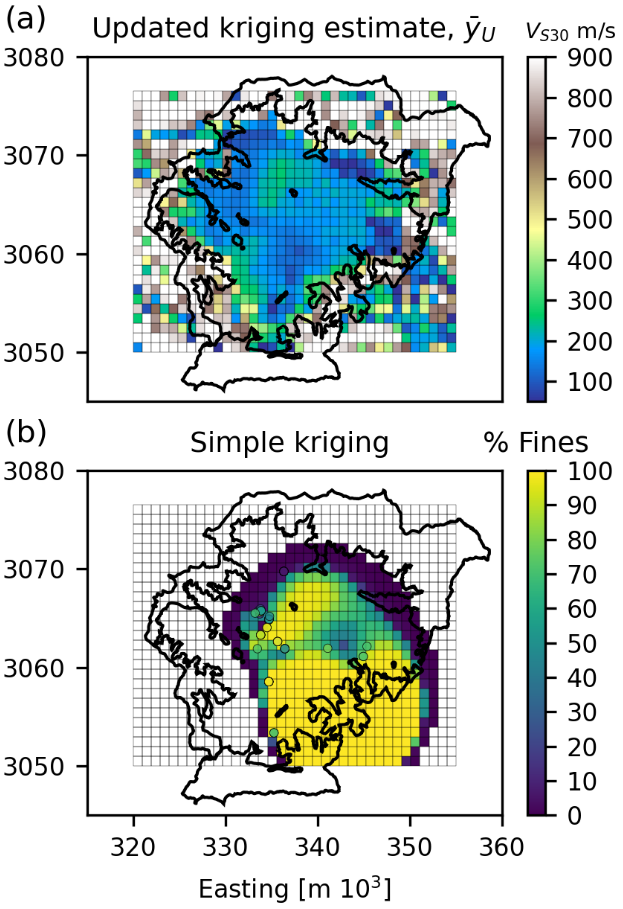

Figure 1: (a) Updated Kriging estimate from Gilder et al. *(2022) compared with (b) Simple Kriging estimate of % fines.*

The *% fines* result also does not follow the trend expected from the geological knowledge of the region, where granular materials dominate the northern portion of the basin (Shrestha *et al.* 1998). This corroborates the authors' previous work, where splitting the correlation of geotechnical parameters vs V_{S30} for grain size did not significantly improve results (Gilder *et al.* 2021).

This is an example of how a suitable candidate as secondary predictor such as *% fines* cannot be then used further in the framework due to insufficient geographical distribution affecting its informative value in the framework as in this case for V_{S30}.

4.2. Spatial prediction of N using the framework

SPT-N or *N* is the blow count measured during a Standard Penetration Test (denoted as *SPT-N* in the following). In the context of earthquake engineering, various authors have explored indirect prediction methods involving the correlation of *Vs* with this geotechnical parameter (*e.g.*, Otha & Goto 1978). In this section, the Kathmandu database (Gilder *et al.* 2020) is used as a source of *SPT-N* data to understand its regional distribution making use of the Gilder *et al.* (2022) framework developed for V_{S30}.

In Figure 2a, the primary data points for *SPT-N* are averaged across the soil column at each location. The data over intervals (up to the maximum depth of the data of 35m) are compared to understand if this is a good approximation. In Figure 3, four histograms are presented, produced of averaged *SPT-N* values for intervals of 5m, 10m, 20m and >20m (data does not include the preceding interval). The frequency of averaged *SPT-N* values does not change significantly until the interval of >20m is reached. It is concluded that taking an average of the entire soil column will not affect the resulting distribution, and it can be accepted as an approach for this dataset.

The Simple Kriging estimate for *SPT-N* (Figure 2b) is indicating several zones of low *SPT-N* values (below 5 blows), and the mean value of the dataset is 13. At the edges of the modelling area, higher values of *SPT-N* are observed, and are increasing to 50, guided by point data present at the western side of the Valley. The inner black line represents the Valley boundary between the relatively flat, level inner portions and the outer mountainous areas. The primary data and the Simple Kriging results (Figure 2) are taken through to the Bayesian analysis (*i.e.*, step 5 and 6 presented in section 3).

In the first section of this paper, *% fines* was considered as a possible secondary predictor for

V_{S30}. Here, this variable is considered again in the context of improving the Simple Kriging result for *SPT-N*. Where data are collocated, Figure 4 shows that there is a reasonable correlation between the two parameters. However, for use in the framework, *% fines* is needed at all unknown locations which are to be considered in the estimate, for it to serve as a secondary data set. As seen in the first part of this paper for the case of V_{S30}, the dataset for *% fines*, even when Kriging is undertaken, remains too few to present a reasonable definition across the Kathmandu Valley.

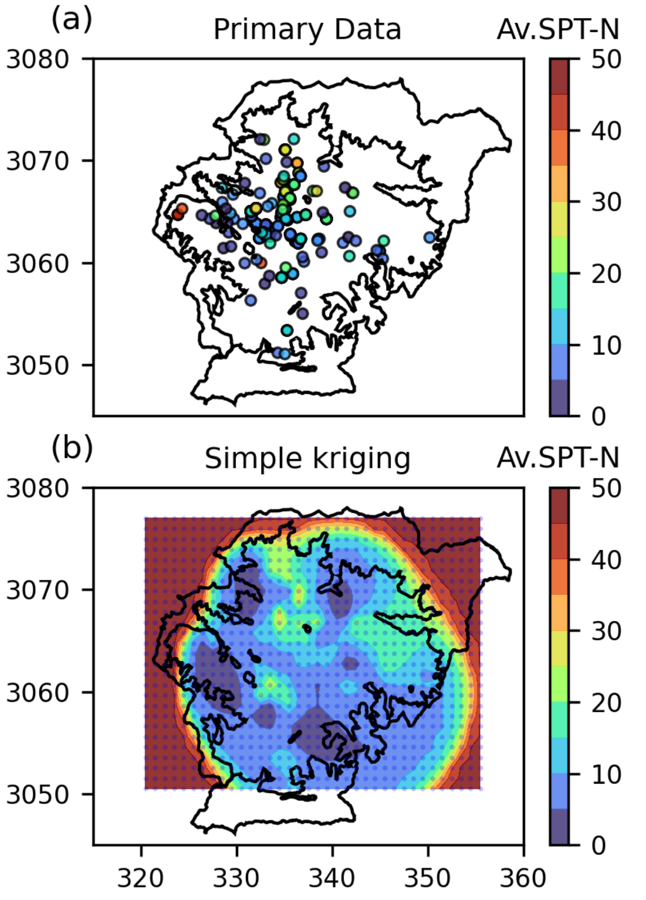

Figure 2: (a) Average SPT-N primary data points (b) Simple Kriging estimate for average SPT-N.

A more robust secondary dataset for the Valley, and a dataset which is commonly available in other regions, is the Digital Elevation Model (DEM). This provides the elevation across the study area in meters above mean sea level (m AMSL). In Figure 5a, the DEM data is shown, and the data ranges between 900m and 2600m AMSL.

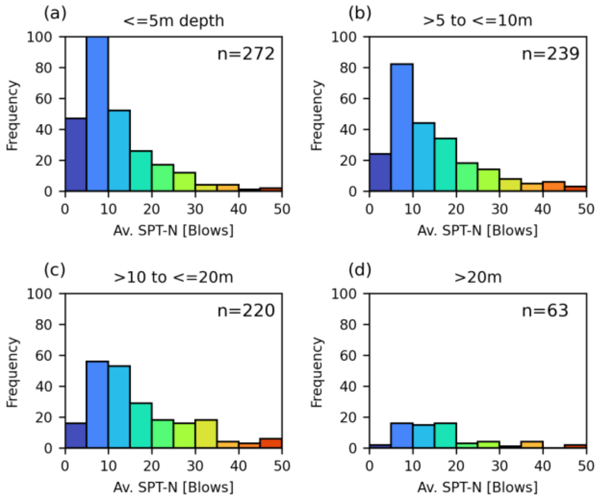

Figure 3: Histograms of average N values quantified at each depth increment (a) <=5m, (b) >5 to <=10m, (c) >10 to <=20m,(d) >20m)

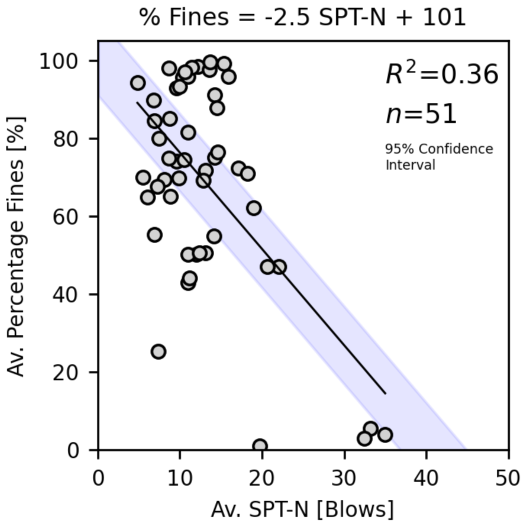

Figure 4: Collocated data for SPT-N and % Fines (n = number of data-points, R^2 = coefficient of determination)

Step 2 of the framework (Gilder *et al.* 2022) is required to ensure that data is not either clustered, causing some concentration of values within the dataset or biased (missing an interval in the dataset altogether). This step can be achieved by inspection of the primary data histogram.

The histograms in Figure 3 indicate that the current *SPT-N* data for the Valley is not biased *i.e.*, the full possible range of data (blows 0 to 50) are represented. However, the dataset may be considered clustered. To understand this concept better and how clustered data might affect the prediction, in Figure 5b, the histogram, presenting the ground level data (m AMSL) at the point of the *SPT-N* measurements is shown to highlight how little the *SPT-N* values are distributed at points of high and low topography in the modelling area. This histogram is compared to the histogram in Figure 5c, which presents all topographic data taken at each square of the DEM (Figure 5a).

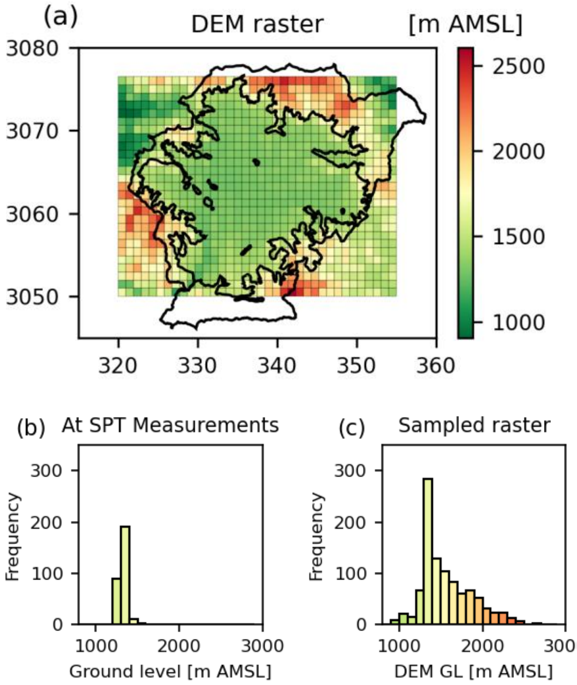

Figure 5: (a) DEM showing the Kathmandu Valley in the centre; (b) Histogram presenting DEM data collocated with primary N data points; (c) Histogram of DEM sampled at the unknown locations to use as secondary data.

Where the DEM indicates a mountainous area, it might be expected to have a high *SPT-N* value, such as >50, which is indicative of rock or highly dense material. When considering the decision between Option A (debiasing) and Option B (declustering) of the framework, the primary data histogram should be consulted in the first instance. Additional geological or judgement-based understanding of the dataset can be gained from other evidence, such as that presented in Figure 5, and this is needed prior to attempting any modelling scenario. Both Figures 3 and 5 have helped to understand that Option B of the framework should be selected in this case and that the primary data distribution can be declustered using Equation 4. As this Kriging method requires the use of Gaussian techniques, all variables are to be normally distributed, which requires a transformation. Once the declustering weights have been defined, the declustered cumulative distribution function (CDF) can be established by weighting the statistical function. This provides a way to transform to the standard normal space and back. In Figure 6, the original CDF of the *N* data is compared to the declustered CDF.

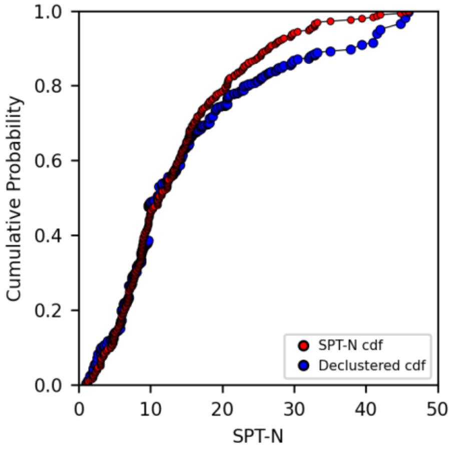

Figure 6: The original CDF of the blow count in SPT data compared to the declustered CDF

Where the primary data points for *SPT-N* (Figure 2a) are concentrated predominantly in the centre of the map, and values are lower (between 5 and 10), these data will be assigned the lowest weights to transfer the weight to areas of the model which represent more sparsely sampled areas. This has the effect of increasing the sample mean from 13 to 15. The declustered CDF in Figure 6 has translated downwards in the upper values of *SPT-N*, indicating that the distribution is now corrected for the predominance of lower values, and this fits with the observation of the higher declustered average. The population is considered to be representative of the entire study area, and the final steps of the framework can be taken. As per the original framework, Equation 1 and 2 are used to produce the Bayesian result. To fully understand the constituent parts of this analysis, further equations are provided by Gilder *et al.* (2022). The main steps are in completing the simple Kriging, so finding the $\bar{y}_P$ for each unknown location (in this case, the result in Figure 2b, but maintaining it in normal space), also defining the secondary variable (*i.e.*, all values at each unknown location in Figure 5a) and converting again to normal space prior to use in the equations. The likelihood is defined between the secondary variable/s and the primary by calculating the Pearson's correlation coefficients (see Gilder *et al.* 2022 for further details). Both these steps result in an accompanying quantification of the estimation variance, parameters $\sigma^2{}_P$ and σ_L^2.

The final calculation gives the updated Kriging map in Figure 7. The result shows where prediction has been determined by the primary data alone (data in the basins) and where the prediction refers to the DEM map (outside of the main Valley extent) being potentially overestimated. This is based on the hypothesis that the mountainous areas might contain rocky outcrops, which would result in a high *SPT-N* values. Where lower topographic levels might be expected to be level and contain sediments, the updated estimate provides an enhanced distribution and further clarification of possible changes which might occur at the extremities of the model. In the centre of the model, where most data is known, the Kriging estimate has maintained the local distribution (compared to the simple Kriging in Figure 2), and so provides a good mix of both detail and combination of judgment.

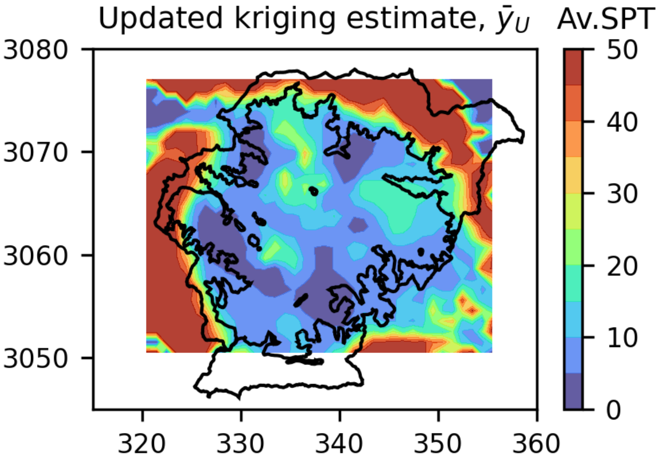

Figure 7: Updated SPT-N Kriging map

5. CONCLUSIONS

The use of a multivariate geo-statistical method to predict *V_{S30}*, which includes steps for debiasing or declustering the data, is particularly relevant in data-scarce regions. The original study by Gilder *et al.* (2022) provided a spatial prediction of *V_{S30}* quantified using multivariate geo-statistics. In this paper, a novel application of the framework to the case study of the Kathmandu Valley is presented. A further estimate for blow count measured during a Standard Penetration Test is presented, attempting the use of different variables as secondary data and finally using a digital elevation model to obtain an updated Kriging map of SPT-N for the Kathmandu Valley. This highlighted a further use of the framework, which can be flexibly implemented in other earthquake-prone regions.

6. REFERENCES

Allen, T. I., & Wald, D. J. (2009). On the use of high-resolution topographic data as a proxy for seismic site conditions (*V_{S30}*). *Bull. Seism. Soc. Am.*, 99(2A), 935–943.

CEN (2004). Eurocode 8: Design of Structures for Earthquake Resistance—Part 1: General Rules, Seismic Actions and Rules for Buildings. European Committee for Standardization (CEN), Brussels.

De Risi, R., F. De Luca, C. E. L. Gilder, R. M. Pokhrel, and P.J. Vardanega (2021). The SAFER geodatabase for the Kathmandu valley: Bayesian Kriging for data-scarce regions, *Earthq. Spectra* 37, no. 2, 1108–1126.

Deutsch, C. V., and A. G. Journel (1998). GSLIB Geo-statistical Software Library and User's Guide, Second ed., Oxford University Press, New York.

Deutsch, C.V, Frykman, P., & Xie, Y.L. (1999). Declustering with Seismic or "Soft" Geological Data. In: *Center for Computational Geostatistics Annual Report Papers*, Report One 1998/1999. University of Alberta.

Deutsch, C.V., & Zanon, S.D. (2007). Direct prediction of reservoir performance with Bayesian updating. *Journal of Canadian Petroleum Technology*, Calgary, Alta: Canadian Institute of Mining and Metallurgy, Petroleum and Natural Gas Division Richardson, Tex.: Society of Petroleum Engineers, 46(2).

Gilder, C.E.L. (2022). Geotechnical data curation and a geostatistical multivariate framework for V_S prediction in data scarce contexts. *Ph.D. thesis*, University of Bristol, Bristol, UK.

Gilder, C.E.L., Pokhrel, R.M., Vardanega, P.J., De Luca, F., De Risi, R., Werner, M.F., Asimaki, D., Maskey, P.N., & Sextos, A. (2020). The SAFER geodatabase for the Kathmandu Valley: geotechnical and geological variability. *Earthq. Spectra*, 36(3), 1549-1569.

Gilder, C.E.L., Vardanega, P.J., Pokhrel, R.M., De Risi, R., De Luca, F. (2021). Assessing Transformation Models Using a Geo-Database of Site Investigation Data for the Kathmandu Valley, Nepal. In: Barla, M. et al. (eds) *Challenges and Innovations in Geomechanics. IACMAG 2021. Lecture Notes in Civil Engineering*, v125: 331-338, Springer, Cham.

Gilder, C.E.L., De Risi, R., De Luca, F., Pokhrel, R.M, & Vardanega, P.J. (2022). Geo-statistical framework for estimation of V_{S30} in data scarce regions, *Bull. Seism. Soc. Am.*, 112(6), 2981–3000.

Mital, U., Ahdi, S., Herrick, J., Iwahashi, J., Savvaidis, A., & Yong, A. (2021). A Probabilistic Framework to Model Distributions of V_{S30}. *Bull. Seism. Soc. Am.*, 111(4), 1677–1692.

Ohta, Y., & Goto, N. (1978). Empirical shear wave velocity equations in terms of characteristic soil indexes. *Earthq Eng Struct Dyn*, 6(2), 167–187.

Shrestha, O.M., Kolrala, A., Karmacharya, S.L., Pradhananga, U.B., Pradhan, P.M., Karmacharya, R., (1998). Engineering and Environmental Geological map of the Kathmandu Valley, Scale 1:50,000. Kathmandu, Nepal: *Department of Mines and Geology*, Lainchaur.

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 V_{S30} from geologic- and terrain-based proxies. *Bull. Seism. Soc. Am.*, 104(6), 2827–2841.

Thompson, E.M., Wald, D.J., & Worden, C.B. (2014). A V_{S30} Map for California with geologic and topographic constraints. *Bull. Seism. Soc. Am.*, 104(5), 2313–2321.

Wackernagel, H. (2010). Multivariate Geostatistics: An introduction with Applications, third edition (completely revised). Springer, Germany.

Wair, B.R., DeJong, J.T., and Shantz, T. (2012). Guidelines for Estimation of Shear Wave Velocity Profiles. PEER Report 2012/08, Pacific Earthquake Engineering Research Centre, University of California, Berkeley, CA, USA.

Wald, D.J., & Allen, T.I. (2007). Topographic Slope as a Proxy for Seismic Site Conditions and Amplification. *Bull. Seism. Soc. Am.*, 97(5), 1379–1395.

Wald, D.J., Worden, C.B., Quitoriano, V., Pankow, K.L. (2006). ShakeMap® Manual, technical manual, users guide, and software guide, available at: http://pubs.usgs.gov/tm/2005/12A01/pdf/508TM12-A1.pdf.

Wills, C.J., & Clahan, K.B. (2006). Developing a map of geologically defined site-condition categories for California. *Bull. Seism. Soc. Am.*, 96(4A), 1483–1501.

Wills, C.J., Gutierrez, C.I., Perez, F.G., & Branum, D.M. (2015). A next generation V_{S30} map for California based on geology and topography. *Bull. Seism. Soc. Am.*, 105(6), 3083–3091.

Yong, A., S. E. Hough, J. Iwahashi, and A. Braverman (2012). A terrain-based site-conditions map of California with implications for the contiguous United States, *Bull. Seism. Soc. Am.*,102(1), 114–128.