https://doi.org/10.32762/eygec.2025.23

# GEOSYN: SYNTHETIC GEOTECHNICAL CROSS-SECTIONS FOR MACHINE LEARNING APPLICATIONS

**Fabian CAMPOS MONTERO[1,2], Eleni SMYRNIOU[1], Bruno ZUADA COELHO[1], Riccardo TAORMINA[3], Philip J VARDON[3]**

**ABSTRACT**

The application of machine learning in geotechnical engineering is often hindered by the scarcity of high-quality, labelled datasets. To address this, we introduce GeoSyn, an open source Python-based tool that generates synthetic geotechnical 2D cross-sections, allowing users to define layer size and number, geotechnical properties and anisotropy with random fields, and boundary conditions. The generated data provides an effective solution for the development and training of ML applications in geotechnics. We demonstrate the tool's utility through two applications. First, we show how a conditional Generative Adversarial Network, trained with synthetic data from GeoSyn, can interpret geotechnical schematisations from Cone Penetration Tests. Second, we explore how Deep Reinforcement Learning can be used to optimise the placement of subsequent in-situ surveys based on prior results. These examples illustrate how GeoSyn enables the development of ML models by leveraging large, flexible datasets to support decision-making in geotechnical engineering.

**Keywords:** synthetic data, machine learning, open source, cone penetration test (CPT), cross-section, random fields.

## INTRODUCTION

The integration of machine learning (ML) in geotechnical engineering is growing rapidly, as shown by Liu et al. (2024) and Yaghaoubi et al. (2024). This growth is driven by ML's ability to detect complex, non-linear relationships and identify patterns overlooked by traditional methods (Alpaydin, 2021).

ML model performance depends on three factors: training data, algorithms, and computational resources (Villalobos et al., 2024). Among these, data availability is often the main constraint. Unlike computer vision or NLP, which rely on large labelled datasets, geotechnical engineering is limited by fragmented data, making model training and generalisation challenging. Data collection is inherently difficult due to the subsurface's heterogeneous, anisotropic, and irregular nature (Phoon & Zhang, 2023).

Geotechnical datasets are also highly contextual, depending on local geology, site conditions, and engineering practices (Phoon, Ching, & Shuku, 2022). These factors hinder ML models from generalising beyond the sites on which they were trained. Expert judgment remains critical, as engineers interpret incomplete datasets and apply domain knowledge to compensate for missing information, a process not yet well integrated into ML workflows (Phoon, Ching, & Cao, 2022).

To address this, we present GeoSyn, an open-source Python tool for generating synthetic 2D geotechnical cross-sections. Users can define parameters such as the number of layers, geotechnical properties, and anisotropy using random fields, creating diverse datasets for ML training.

This paper is structured as follows: Section 2 explains GeoSyn's rationale and assumptions. Section 3 details geometry generation. Section 4 presents two ML applications: a conditional Generative Adversarial Network for geotechnical interpretation and a Deep Reinforcement Learning model for optimising in-situ survey placement. Section 5 discusses results and limitations, followed by conclusions in Section 6.

## SYNTHETIC SUBSURFACE MODELLING FRAMEWORK

The goal of synthetic geotechnical cross-sections is to create diverse, realistic subsurface representations for training and validating ML models. Any scalar geotechnical indicator can be used for schematisation.

Here, we focus on the Soil Behaviour Type Index (IC) proposed by Robertson (1990), derived from Cone Penetration Tests (CPT), widely used in practice, especially in the Netherlands. However, the methodology is not limited to this parameter

1 *Deltares, Department of Geo-Engineering, Delft, The Netherlands*
2 *Corresponding author: fabian.campos@deltares.nl (F.A. Campos Montero)*
3 *Delft University of Technology, Faculty of Civil Engineering and Geosciences, Delft, The Netherlands*

and can also be applied to others, such as the Unified Soil Classification System, undrained shear strength, permeability, or thermal conductivity. As long as the property can be expressed as a 2D field, the approach remains valid.

**Size and format**

Each synthetic cross-section is a 2D array (512 × 32 pixels), with depth on the vertical axis and horizontal distance across the site. This format balances visual clarity, resolution, and computational efficiency for ML applications. Users can adjust the array size as needed without altering the methodology.

**Geotechnical assumptions**

The subsurface is represented as a layered system. The number of layers varies per user-defined parameters. While examples in this paper use up to five layers, the method supports any number, offering flexibility for a wide range of geological scenarios. Layers differ in thickness and geometry, including undulating boundaries, indentations, lenses, and in-filled gullies, mimicking natural depositional and erosional processes.

A flat ground surface is assumed, providing a consistent top boundary and simplifying alignment across synthetic cases. Within each layer, materials are spatially variable and anisotropic. Rather than uniform conditions, target properties are modelled as random fields, introducing heterogeneity. This allows lateral continuity and vertical variation, reflecting real-world soil behaviour and improving dataset realism.

## CROSS-SECTION GENERATION METHODOLOGY

Synthetic cross-section generation involves two steps: first, creating the geometric layer structure, and second, assigning material properties as scalars or spatially variable fields (Figure 1).

**User defined parameters**

GeoSyn allows users to define key parameters directly within the Python code. These parameters govern how each synthetic cross-section is generated and offer flexibility to simulate a wide variety of subsurface conditions. Users can control:

- Number of layers to be created (up to any desired maximum),
- Boundary geometry through amplitude, wavelength, vertical shift, and phase shift of sine or cosine functions,
- Assignment of geotechnical properties, either by filling layers with single scalar values (e.g., for soil classes) or using 2D random fields,
- Material definitions, including distributions of IC values, spatial variability, and correlation lengths for each soil type,
- Anisotropy settings for horizontal and vertical correlation lengths in the random fields,
- Order of layer filling, which can follow a fixed sequence, a random shuffle, or a hybrid approach.

**Layer boundary geometry**

Layer boundaries are generated procedurally using sine or cosine functions with randomly sampled parameters: amplitude (vertical relief), period (spacing), vertical shift (overall depth), and phase shift (horizontal translation).

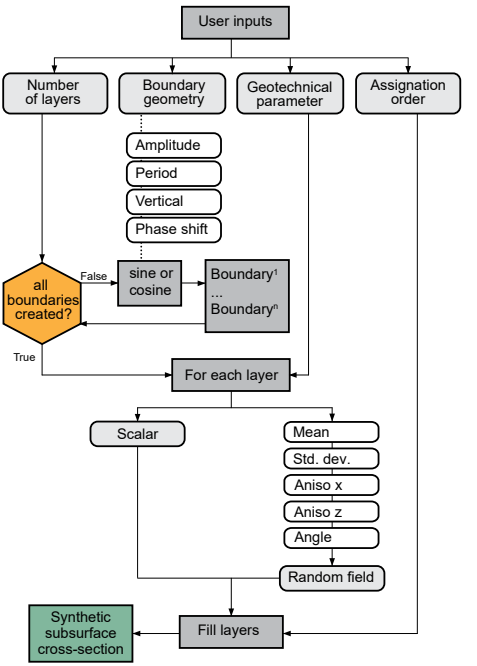


**Figure 1** *Schematic representation of the synthetic cross-section generation process*

Amplitudes and wavelengths are sampled from PERT distributions, enabling likely (most probable) values while permitting variation. Vertical and phase shifts use uniform distributions, ensuring all allowable depth and position values are equally probable.

The layering process is sequential. Interactions between boundary curves, especially with high amplitudes or large phase shifts, may result in fewer distinct layers. These cases produce enclosed shapes like lenses, channels, or indentations, enhancing geological realism.

Once boundaries are defined, the space between them is discretised into pixel-level regions representing individual layers. These regions provide spatial domains for assigning properties in the next stage.

**Property assignment and spatial variability**

Geotechnical parameters are assigned to each layer. For example, five representative soil categories were defined, each characterised by statistical IC-value distributions from literature. Sand layers have lower IC-values (1.3–2.0), while clay and organic layers exhibit higher values (3.0–4.0).

Within each layer, GeoSyn generates a 2D Gaussian random field via GSTools (Müller et al., 2022) to simulate pixel-level heterogeneity. Anisotropy is introduced by defining horizontal and vertical correlation lengths and rotation angles, reflecting depositional processes and typical soil behaviour.

The final output is a continuous IC-value field for each layer, merged into a single raster image representing the synthetic cross-section ( Figure 2).

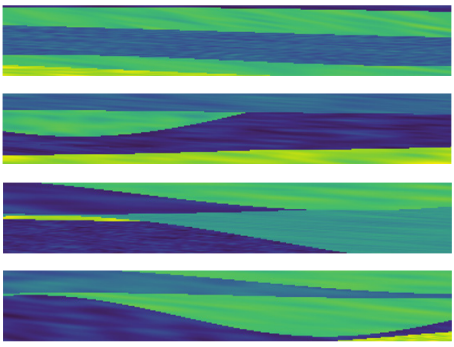

**Figure 2** *Synthetic models simulating the anisotropic spatially variable IC-values.*

## APPLICATIONS

To demonstrate GeoSyn's utility, we present two applications: a conditional Generative Adversarial Network (cGAN) for stratigraphic reconstruction and a Deep Reinforcement Learning (DRL) model for optimising in-situ test planning. Both rely on synthetic datasets generated with GeoSyn.

**Shared synthetic database**

Both applications use a synthetic dataset of 24,000 cross-sections (512×32 pixels) openly available on Zenodo (Campos Montero, 2024). These images represent IC-values and include a range of geological features from simple sub-horizontal layering to complex lenses, indentations, and buried channels. The dataset reflects deltaic conditions typical in the Netherlands and exposes ML models to diverse learning scenarios.

**Geotechnical schematisation using conditional GANs.**

GeoSyn was used to train SchemaGAN, a cGAN that infers complete subsurface schematisations from sparse CPT-like data (Figure 4). The model inputs images with <1% of original data, simulating real-world CPT tests, and reconstructs full cross-sections with layer geometry and intra-layer variability.

GeoSyn provided ground truth images for training, validation, and testing. Sparse inputs were created by removing 99% of data from each synthetic image and leaving a few vertical CPT-like columns. These sparse images, paired with their full-resolution counterparts, enabled adversarial training: the discriminator distinguished real from generated cross-sections, while the generator redined outputs to match GeoSyn ground truths.

SchemaGAN generalised beyond simple geometries, reproducing complex subsurface features. It outperformed traditional interpolation methods in both synthetic and real-case tests, capturing smooth transitions, irregular interfaces, and internal heterogeneity. Full methodology and evaluation are detailed in Campos Montero et al. (2025).

**Optimisation of site investigation with reinforcement learning**

In the second application, GeoSyn trained a DRL agent to optimise CPT placement. The goal was to minimise tests while maintaining accurate subsurface reconstruction. Details of the model and architecture are provided in Zuada Coelho et al. (2025).

The RL environment used full-resolution GeoSyn images. In each episode, the agent selected in-situ test locations and received rewards balancing efficiency (fewer tests) and accuracy, measured by RMSE between predicted and ground truth profiles. Predictions used Inverse Distance Weighting (IDW) based on selected CPTs.

Exposure to diverse synthetic profiles with complex transitions (e.g., lenses, sharp boundaries) enabled the agent to learn where denser testing was needed. The DRL agent consistently outperformed fixed-spacing strategies, adapting test placement to local soil complexity and achieving better accuracy with fewer tests.

## RESULTS AND DISCUSSION

The two ML applications highlight GeoSyn's versatility in enabling data-driven approaches in geotechnical engineering. Both methods, though addressing different tasks (interpretation and planning), benefited from training on the same synthetic dataset and showed strong performance.

SchemaGAN achieved robust results in both synthetic validation and real-case applications. It consistently outperformed conventional interpolation methods, capturing subtle transitions, irregular boundaries, and internal variability. This underscores the value of training ML models on diverse subsurface conditions, as enabled by the GeoSyn dataset.

The DRL agent for CPT placement also benefited from varied synthetic profiles. Instead of uniform spacing, the agent adapted test locations to soil complexity, achieving accurate reconstructions with fewer tests. The largest gains occurred in complex profiles with lenses, steep boundaries, and abrupt transitions. This demonstrates how training on geologically plausible schematisations supports both interpretation and investigation planning.

A key advantage in both cases was access to fully labelled data, rare in real geotechnical practice. For SchemaGAN, this enabled adversarial training with direct comparisons to ground truth. For the DRL agent, full IC-fields allowed precise RMSE calculations, providing reliable reward signals for learning. Without synthetic data, such models would have been far harder to develop and evaluate.

Notably, the dataset was not based on specific site investigations or calibrated with project statistics. Instead, it represented general deltaic conditions, covering a wide range of plausible scenarios. Despite this generality, both ML applications performed well, even on real-case data. This suggests general-purpose synthetic datasets can effectively support early ML development, with refinements added later for specific geological settings.

These results show synthetic data's potential to address key challenges in geotechnical ML: data scarcity, lack of standardisation, and validation difficulties. GeoSyn enables robust, reproducible, and scalable ML workflows.

Some limitations remain. GeoSyn assumes planar top surfaces, uses a fixed grid resolution, and focuses on scalar IC-values, which may not fully capture real subsurface complexity. Its outputs depend on how representative the user-defined parameters are. For larger or more heterogeneous sites, multiple parameter sets may be needed to reflect distinct geological zones.

While both ML applications generalised well, applying models trained on synthetic data to real projects still requires caution. Broader validation against diverse field datasets is essential to build confidence in these approaches.

GeoSyn's flexibility also allows future extensions. Users can model different parameters (e.g., undrained shear strength, permeability), modify random field structures, or include site-specific geostatistical constraints. Its open-source design supports collaborative use in research and practice

.

## CONCLUSIONS

This paper introduced GeoSyn, an open-source tool designed to generate synthetic geotechnical cross-sections for machine learning applications. By providing control over layering, spatial variability, and material properties, the tool enables the creation of realistic and diverse datasets that address a fundamental bottleneck in data-driven geotechnics: the scarcity of high-quality, labeled data.

We demonstrated its utility through two distinct applications: subsurface reconstruction with conditional GANs and in-situ test placement optimisation using reinforcement learning. In both cases, the synthetic data enabled the training of models that would otherwise be difficult to develop using only real-world data.

GeoSyn offers a flexible and reproducible framework to support the development and testing of machine learning methods in geotechnical engineering. Its public release aims to facilitate further research and collaboration in data-centric approaches to subsurface characterisation. By enabling rapid model training using synthetic data, GeoSyn can support early-stage site assessments, guide the design of investigation campaigns, or serve as a testing ground for new ML methods before applying them to costly real-world data.

## DATA ACCESIBILITY

The GeoSyn tool is open access and free to explore at https://github.com/fabcamo/GeoSyn. The data based used in the given examples can be accessed at https://zenodo.org/records/13143431.

## REFERENCES

Alpaydin, E. (2021). Machine learning (Revised and updated edition). The MIT Press.

Campos Montero, F. (2024). SchemaGAN: Dataset and Pre-trained Model for Subsoil Schematization (Version 1). Zenodo. https://doi.org/10.5281/ZENODO.13143430

Campos Montero, F. A., Zuada Coelho, B., Smyrniou, E., Taormina, R., & Vardon, P. J. (2025). SchemaGAN: A conditional Generative Adversarial Network for geotechnical subsurface schematisation. Computers and Geotechnics, 183, 107177. https://doi.org/10.1016/j.compgeo.2025.107177

Ching, J., & Phoon, K.-K. (2020). Constructing a Site-Specific Multivariate Probability Distribution Using Sparse, Incomplete, and Spatially Variable (MUSIC-X) Data. Journal of Engineering Mechanics, 146(7), 04020061. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001779

Liu, H., Su, H., Sun, L., & Dias-da-Costa, D. (2024). State-of-the-art review on the use of AI-enhanced computational mechanics in geotechnical engineering. Artificial Intelligence Review, 57(8), 196. https://doi.org/10.1007/s10462-024-10836-w

Müller, S., Schüler, L., Zech, A., & Heße, F. (2022). GSTools v1.3: A toolbox for geostatistical modelling in Python. Geoscientific Model Development, 15(7), 3161-3182. https://doi.org/10.5194/gmd-15-3161-2022

Phoon, K.-K., Ching, J., & Cao, Z. (2022). Unpacking data-centric geotechnics. Underground Space, 7(6), 967-989. https://doi.org/10.1016/j.undsp.2022.04.001

Phoon, K.-K., Ching, J., & Shuku, T. (2022). Challenges in data-driven site characterization. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 16(1), 114-126. https://doi.org/10.1080/17499518.2021.1896005

Phoon, K.-K., & Zhang, W. (2023). Future of machine learning in geotechnics. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 17(1), 7-22. https://doi.org/10.1080/17499518.2022.2087884

Robertson, P. K. (1990). Soil classification using the cone penetration test. Canadian Geotechnical Journal, 27(1), 151-158. https://doi.org/10.1139/t90-014

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data (No. arXiv:2211.04325). arXiv. https://doi.org/10.48550/arXiv.2211.04325

Yaghoubi, E., Yaghoubi, E., Khamees, A., & Vakili, A. H. (2024). A systematic review and meta-analysis of artificial neural network, machine learning, deep learning, and ensemble learning approaches in field of geotechnical engineering. Neural Computing and Applications, 36(21), 12655-12699. https://doi.org/10.1007/s00521-024-09893-7

Zuada Coelho, B., Smyrniou, E., & Campos Montero, F. A. (2025). In-situ site investigations with Deep Reinforcement Learning. Geodata and AI (Submitted).

This page was intentionally left blank