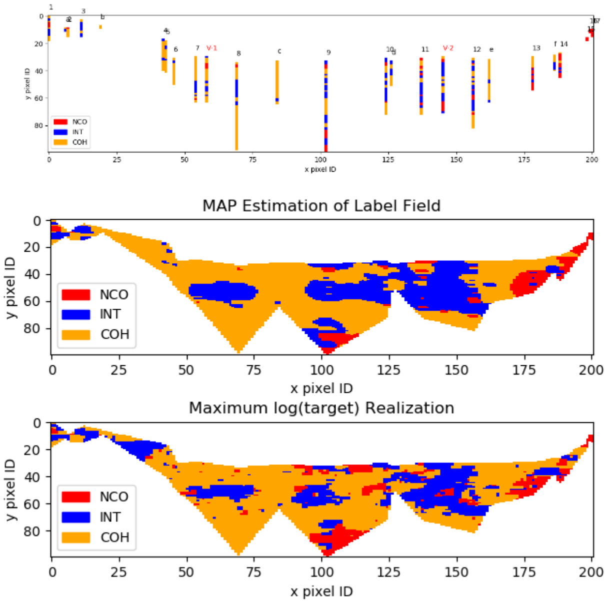
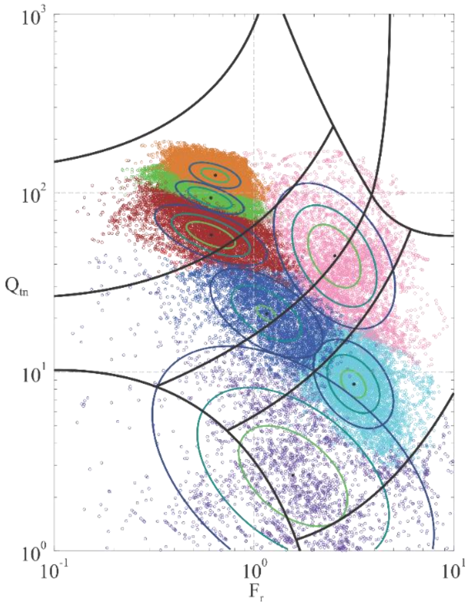
Study of AI Based Methods for Characterization of Geotechnical Site Investigation Data





Prepared by:
Hui Wang, Xiangrong Wang, Robert Liang

Prepared for:
The Ohio Department of Transportation,
Office of Statewide Planning & Research

State Job Number 135785
Task# 6

February 2020

Final Technical Report

U.S. Department of Transportation
Federal Highway Administration

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA/OH-2020-3		

4. Title and Subtitle	5. Report Date
Study of AI Based Methods for Characterization of Geotechnical Site Investigation Data	**February 2020**
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Hui Wang, Xiangrong Wang, Robert Liang	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
The University of Dayton 300 College Park Drive Dayton, OH 45469-0243	
	11. Contract or Grant No.
	SJN 135785 Task#6
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Ohio Department of Transportation 1980 West Broad Street Columbus, Ohio 43223	**Final Technical report May. 2019 – Jan. 2019**
	14. Sponsoring Agency Code

15. Supplementary Notes

Prepared in cooperation with the Ohio Department of Transportation (ODOT) and the U.S. Department of Transportation, Federal Highway Administration

16. Abstract

Due to the inadequate knowledge of the soil forming histories and/or human activities, the subsurface soil layers are difficult to ascertain. Subsurface uncertainty and its influence on geotechnical design have long been a challenge facing practitioners. Recently, the ASCE Geo-institute has developed the Data Interchange for Geotechnical and Geoenvironmental Specialists (DIGGS), which is a standard schema for transferring geotechnical data between multiple organizations. It paves the way of sharing and unifying datasets and forms a structural database for further data-driven modeling and analysis. The Office of Geotechnical Engineering at ODOT (OGE) is taking a national leading role in supporting the development efforts of DIGGS and hence make this project possible. In this study, site investigation data in DIGGS format and archived format are jointly processed. An innovative technique developed by the research team has been further improved for better application in real-world projects. Bayesian machine learning is integrated with Markov random field models to infer and simulate subsurface models and geospatial data with quantified uncertainty. Spatial heterogeneity and statistical characteristics are modeled in terms of statistical and spatial patterns. These patterns serve as a basis to provide a synthesized interpretation of the soil profiles with uncertainty quantified. Four (4) validation projects have been performed in this report and the results are well documented. Summary and recommendations for future work are also provided. A short introduction of the key concepts behind this technique, and pathway for converting the existing program into a ready for implementation web-based program for potential ODOT usages are provided in the appendices.

17. Keywords	18. Distribution Statement
DIGGS; subsurface modeling; Bayesian machine learning; uncertainty quantification; Markov random field	**No restrictions. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161**

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	**Unclassified**		

Form DOT F 1700.7 (8-72) **Reproduction of completed pages authorized**

Study of AI Based Methods for Characterization of Geotechnical Site Investigation Data

Prepared by:

Hui Wang, Xiangrong Wang, Robert Liang
The University of Dayton
300 College Park Drive
Dayton, OH 45469-0243

February 2020

Prepared in cooperation with the Ohio Department of Transportation
and the U.S. Department of Transportation, Federal Highway Administration

The contents of this report reflect the views of the author(s) who is (are) responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Ohio Department of Transportation or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

Acknowledgments

The research work conducted in this report was the result of a team effort involving the University of Dayton research team and the members from ODOT Office of Geotechnical Engineering. The authors acknowledge Chris Merklin, Steve Taliaferro, Paul Painter, and Alex Dettloff of the Ohio Department of Transportation Office of Geotechnical Engineering for their efforts in providing comments and suggestions during the investigation, providing the data used in the validation analysis, and reviewing the draft final technical report. The authors are grateful for their diligent efforts and contributing their practical knowledge.

Table of Contents

1 Introduction

Acquiring accurate site-specific soil and rock information is a crucial and essential step for planning and design of any geotechnical construction project. However, subsurface soil/rock layers are natural forming materials associated with inherent heterogeneity and randomness. Therefore, the design and construction of a geotechnical system that is either embedded in or founded on the subsurface soils need to take into account these spatial variations of soil/rock layering and the engineering properties of each identified layer. Due to the nonspecific knowledge of the soil forming histories and/or other prior geological and human activities, the subsurface information at a project site can be difficult to ascertain. Drilling and sampling to obtain borehole logs, together with various in-situ testing, are usually performed by the transportation agencies and/or geotechnical consultants for determining subsurface soil and rock profiles and their associated engineering properties. However, in practice only a limited number of borehole logs and in-situ soundings are conducted for a given project, partially due to the limited budget and the tight project schedule. As a result, the geological and geotechnical information only can be probed at sparsely geographically distributed locations; whereas, subsurface information at other locations may have to be inferred based on available information either from archived data or planned site investigation data. The nonspecific knowledge of the formation process of the geological bodies, together with the insufficient number of borehole logs and in-situ test results, leads to significant uncertainty in the inferred subsurface model. It is fair to say that the issues resulted from the subsurface uncertainty and the influence of such uncertainty on the geotechnical design have long been a challenge facing practitioners. To be more specific, these challenges can be summarized below.

- Available site investigation data is from multiple sources with a variety of degree of accuracy, credibility, and resolution (e.g., borehole drilling and soil sampling vs. in-situ tests vs. geophysical measurement, or historical archived vs. current investigation), hence there is a challenge for performing consistent and rationale data fusion;

- Available subsurface information is geographically distributed and generally sparse, thus requiring rationale interpretation methods;

- There is no methodology to allow for engineers to assess the level of confidence once he or she has developed the interpreted soil/rock layering information and the associated soil/rock properties for subsequent design and construction purpose;

- Engineers spend enormous amount of time to complete interpretation and presentation of the subsurface models, while interpretation often relies on subjective engineering judgement and engineers' preference for simplification;

- Current interpretation methods at best yield a deterministic model where quantitative assessment of uncertainties (or confidence level) of such model is lacking and the effects of such uncertainties on the subsequent engineering analysis/design of geotechnical systems cannot be considered at this stage.

The research team at the University of Dayton has dedicated significant amount of efforts in developing transformational methodologies to overcome these aforementioned challenges, while recognizing early on that digital data of geotechnical investigation information will become more widely available. This transformation into digital data era has been propelled by adoption of a good practice of geotechnical data management (GDM), as it has been ongoing at ODOT, and the emergence of common agreed data formats for geotechnical data. For standardized geotechnical data format, the UK and various parts of the world have reached agreed upon standards, AGS (Association of Geotechnical and Geo-environmental Specialists) (Walthall and Palmer 2006). In the United States, the DIGGS (Data Interchange for Geotechnical and Geoenvironmental Specialists) (Weaver et al. 2008) format is starting to emerge as the preferred format. Both formats enable the transfer of geotechnical and geo-environmental data within and between organizations. When utilizing data interchange standards, compiling geotechnical data requires importing the data from the data interchange file into the chosen geotechnical data management system.

The Office of Geotechnical Engineering at ODOT (OGE) is taking a national leading role in supporting the development efforts of DIGGS. The value of adopting such practice in the industry could be further enhanced, if a practical and user-friendly computational tool can be afforded to the geotechnical practicing industry for addressing the above-mentioned challenges associated with current geotechnical site investigation practices. As geotechnical site investigation data in ODOT will be DIGGS compliant, it is now an opportune moment to develop an integrated computational subsurface interpretation and modeling tool to truly transform geotechnical site investigation and interpretation into a digital world.

Figure 1 depicts the overall framework of the AI based methodology developed by the research team. In brief, it can be separated into four different modules (Module 1-4 in Fig. 1). The mathematical foundation of the developed methodologies is based on the following knowledge base and principles:

- Geo-statistics and Markov random field theory;
- Unsupervised machine learning (Gaussian mixture model);
- Bayesian inferential framework;
- Stochastic simulation techniques;
- Information theory.

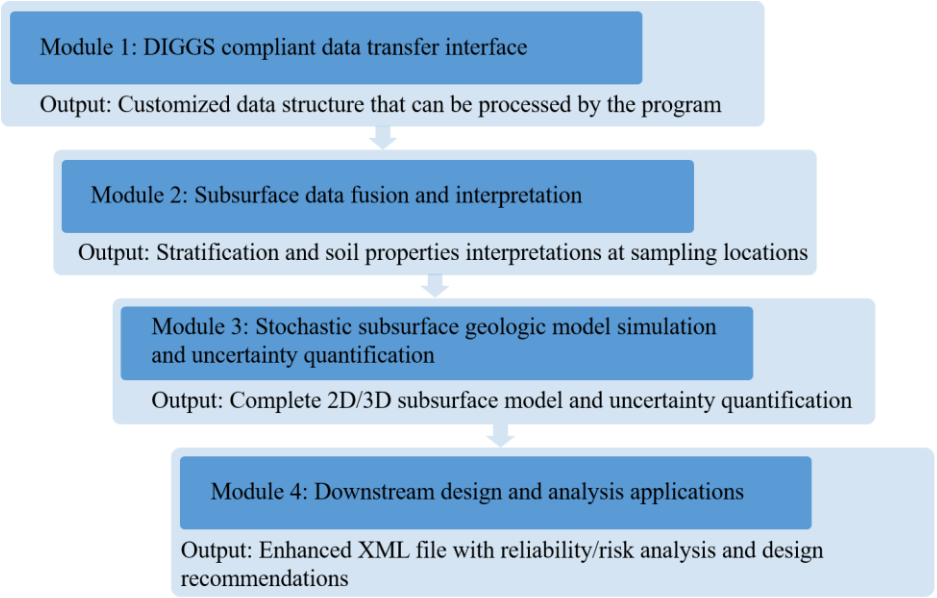

Module 1: DIGGS compliant data transfer interface
Output: Customized data structure that can be processed by the program
Module 2: Subsurface data fusion and interpretation
Output: Stratification and soil properties interpretations at sampling locations
Module 3: Stochastic subsurface geologic model simulation and uncertainty quantification
Output: Complete 2D/3D subsurface model and uncertainty quantification
Module 4: Downstream design and analysis applications
Output: Enhanced XML file with reliability/risk analysis and design recommendations

Figure 1. The overall framework of the developed computational modeling/simulation techniques for geotechnical site characterization

The mathematical foundation, theories, and examples of applications have been published in peer-reviewed high impact journals. A list of publication related to the developed computational algorithms is provided in the references. The developed method is still evolving and functions and performance are continuously improved. Therefore, this project is considered as an initial effort for establishing a robust and practical workflow for AI based subsurface modeling and uncertainty quantification. Ultimately, the developed methodologies can achieve the following two major functions in the future:

- Fuse multiple datasets from different site investigation methods (borehole drilling and sampling, in-situ test and geophysical measurements) and provide synthesized interpretation of the soil/rock profiles with the statistical interpretation of soil/rock properties of each layer based on stochastic simulation, AI, and ML techniques;

- Enhance the visualization of the interpreted subsurface models with added information, such as a measure of confidence level and identification of locations where additional borehole logs and/or CPT sounding could be performed to improve the confidence level of the interpreted subsurface geotechnical models.

1.1 Objectives of this Study

The objectives of the current research can be summarized as follows:

- Study the DIGGS XML based borehole log reporting format and modify the current computer program to read this data as input;

- Work with OGE to perform multiple case studies using ODOT project site investigation data to gain in depth understanding of potential areas for further improvement of the computational algorithms

- Provide a white paper regarding the feasible road map to make the developed program a verified, stand-alone, user friendly, and implementable web-based computational tool that meet ODOT's requirements and application needs.

1.2 Scope of Work

The research work can be divided into three main groups:

Task group 1: Modify the current program to be able to read DIGGS compliant XML-based schema

Task group 2: Perform multiple case studies using benchmark datasets (geotechnical investigation data of several project sites) provided by ODOT. Illustrate and discuss the validity of the essential techniques in the computational program

Task group 3: Conduct in-depth analysis and provide a road map for further work to make the current research grade computer program an implementation-ready, web-based computational software.

2 Python Application Interface (py-API) for processing data with DIGGS schema

2.1 General introduction of DIGGS

DIGGS (Data Interchange for Geotechnical and Geoenvironmental Specialists) is a standard format for the electronic transfer of geotechnical and geoenvironmental data. DIGGS is software neutral and non-commercial. DIGGS can be used for transfer of all geotechnical and geoenvironmental data throughout all project stages, thus offering enormous advantages in terms of workflow efficiency, data accuracy and validity, records retention and management, and consequently cost savings. These features provide an open platform for data documentation, exchange, and processing.

2.1.1 DIGGSML Schema

The DIGGS schemas are Open Geospatial Consortium (OGC) *Geography Markup Language (GML) application schemas* meaning that all schema constructs must derive from GML elements and types, and follow GML's Object/property model, which govern how schema elements and Extensible Markup Language (XML) instance documents are constructed. GML is an XML application that provides a grammar and base vocabulary for describing geo-referenced geotechnical and geoenvironmental data. GML was developed in order to provide a standard means of representing information about geospatial features-their properties, interrelationships, and so on.

Features describe real world entities and are the fundamental objects in GML. Features can be concrete and tangible, such as boreholes and trench walls, or abstract and conceptual, such as projects and jurisdictional boundaries. GML features are described in terms of their properties, which can represent spatial and temporal characteristics or associations with other features. For instance, GML can describe the location, shape, and extent of geographic objects as well as properties such as color, speed, and density, some of which may depend on time. As it is impossible to describe all features for all application domains and predict their usage a priori, the GML core schemas do not fix definitions of specific implementation of feature types such as a trial pits or layer systems. Rather, specific features and properties are defined in GML Application Schemas, which are created by user communities such as DIGGS. So, DIGGS defines the appropriate GML elements and applications used in the delivered schema as applied to Geotechnical and Geoenvironmental engineering.

GML provides a base of common geographic and geometric constructs (e.g. the Abstract Feature model, Points, Line Strings, and Polygons) that can be shared and reused by GML Application Schemas. In turn, the GML constructs are built upon XML constructs such as elements, attributes, types, data types (e.g. integers, strings, dates), international language support, etc. By building on successful existing web technologies, the DIGGS GML Application Schemas can leverage a whole world of GML and XML Tools.

2.1.2 DIGGS Objects

The DIGGS schema contains elements in the form of Objects and Properties. An Object represents a feature (e.g. Borehole, sample, etc.) and then properties about that object (e.g. diameter, height, density, etc.)

Features are the primary *objects* in DIGGS. They are named entities comprised of descriptive properties. Non-feature objects also exist and are structurally the same as features; but, typically are not shared out of context with the features they are associated with. In DIGGS, objects appear as nested complex property values of features (a complex property element is one that contains child elements), e.g. a polygon representation of a trench wall's surface extent. A layer system defining soil descriptions is an example of a DIGGS feature, whereas the individual layers contained within a layer system are just objects that wouldn't be shared outside of the context of the layer system. *Metadata objects* are used to describe contextual information about features or other objects.

2.1.3 DIGGS Properties

Properties are simply child elements of a feature or object. For example, a numeric result of a test is a property of the test feature. Figure 1 illustrates properties as direct children of a Borehole feature.

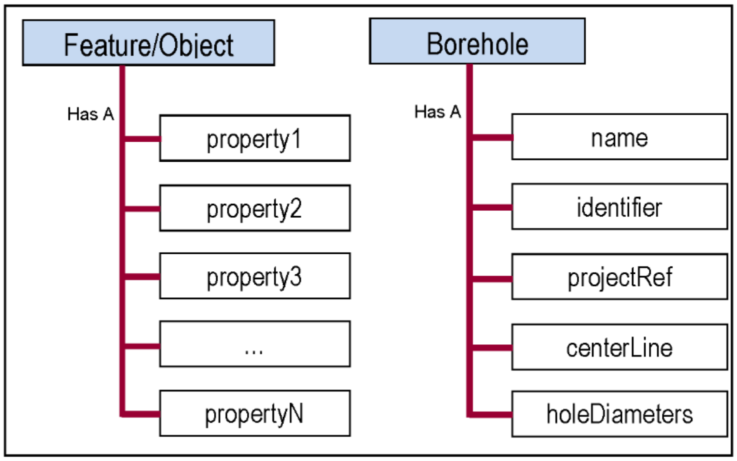


Figure 2. A DIGGS Feature or Object is described by its property children

Figure 2 reveals a GML syntactic convention used to distinguish between Objects and properties; element and type names representing Objects are written in UpperCamelCase and the property names are written in lowerCamelCase.

2.2 Data structure of DIGGS complete .xml file

2.2.1 DIGGS Data storage

Instances of the schema that contain actual data can be created and stored anywhere, online or offline, but were designed for sharing over the web. Data repositories are maintained by DIGGS users and can be read by applications on mobile devices, desktop workstations, or computer servers from various data stores:

File directories – accessible online as public or private web pages or offline in local file directories (e.g. for field work without internet access).

Spatial Databases – accessible online through public or secure web interfaces or offline using a standalone client interface

Data instances can be validated against the official DIGGS schemas online or can be validated by a locally saved/cached copy of the DIGGS schemas.

2.2.2 DIGGS 2.0.a Feature Model

DIGGS 2.0.a defines eight (8) base classes of features (as shown in Figure 3 below) that can be contained as a child under the root DIGGS element. This classification is formalized so that all existing features in DIGGS are categorized by derivation from these base classes. The existing features in DIGGS 2.0.a are the commonly used and requested features by the DIGGS community.

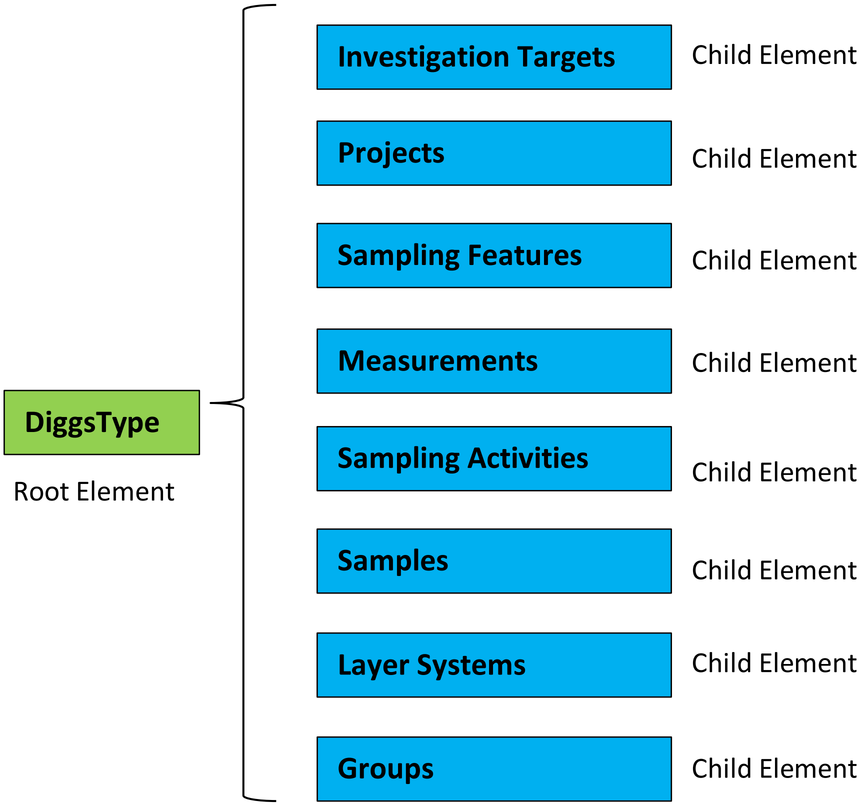


Figure 3. Base Feature Classes in DIGGS 2.0.a

The eight (8) base feature classes are classified by Processes, Entities, and Groups as follows:

- **InvestigationTarget** –target features of interest being sampled/measured [Entity]

- **Project** - business activities that collect, compile, and process information from locations [Process]
- **SamplingFeature** - real world places and constructions (e.g. Boreholes) from which observations are made, samples are collected, or tests are run. [Entity]
- **Measurement** – test readings (in-situ or not) taken from samples collected from sampling features, or created via a sampling activity [Process]
- **SamplingActivity** - the process of sample creation or collection [Process]
- **Sample** - earth material, fluids, or gases collected or created for observation and testing [Entity]
- **LayerSystem** - ordered interval observations or interpretations of earth materials, properties or features at a location [Entity]
- **Group -** collections of projects, locations, samples or groups of these, for the purpose of providing meaningful context to observations and measurements.

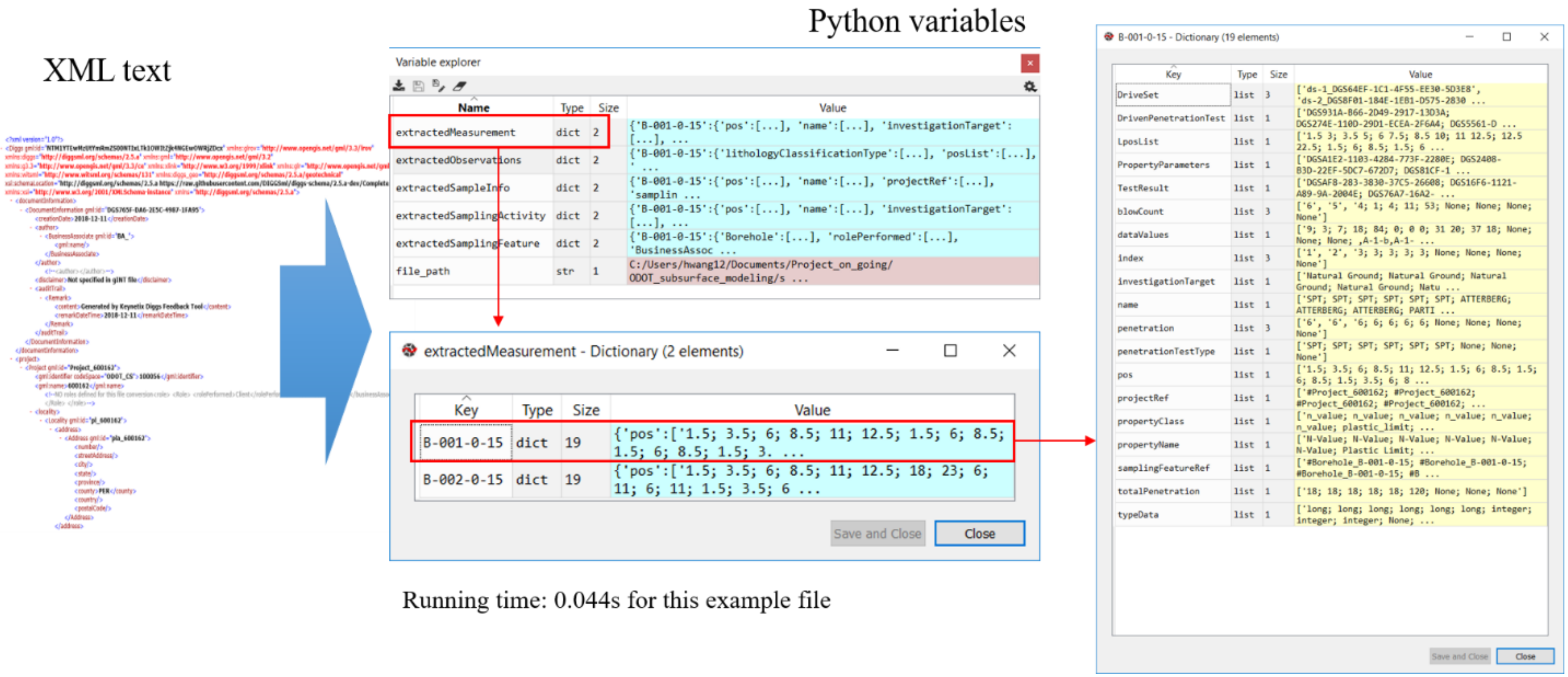
2.2.3 DIGGS Feature Properties and Attributes

DIGGS objects have a number of properties including mandatory and optional. Optional properties of all objects include status, description, and remarks metadata; and all features include additional optional properties including associated file and role metadata objects. Projects, Sampling Features, Samples, Layer Systems, Sensors, and Groups are "named" features. In addition to the identifiers and other properties, they also carry a mandatory name property. Some DIGGS objects are named (i.e. carry a mandatory name property) including some of the layers and all of the Metadata objects.

Objects that need to be referenced within the schema need to have a name. For example, a borehole must have a name, so it can be referred in the schema as to where a sample came from. A sample must have a name so a test can be assigned to the sample. Properties that stay within the hierarchy of the object and need no external reference do not have a mandatory name.

2.3 DIGGS – Python application interface

A python application interface (py-API) has been developed to read and parse the DIGGS .XML files and to construct a customized data structure for the software modules as shown in Figure 1. The developed python interface is based on existing open source python standard library “structured markup processing tools”. The elements (i.e., all the features and their properties) defined in the .xml files are extracted and the corresponding tree structure is converted into native python data structures as shown in Figure 4.



XML text

```
<?xml version="1.0"?>
- <Diggs gml:id="NTM1YTEwMzUYmRmZS00NTIxLTk1OWItZjk4NGEwOWRjZDcx" xmlns:glrov="http://www.opengis.net/gml/3.3/lrov"
xmlns:diggs="http://diggsml.org/schemas/2.5.a" xmlns:gml="http://www.opengis.net/gml/3.2"
xmlns:g3.3="http://www.opengis.net/gml/3.3/ce" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:glr="http://www.opengis.net/gml
xmlns:witsml="http://www.witsml.org/schemas/131" xmlns:diggs_geo="http://diggsml.org/schemas/2.5.a/geotechnical"
xsi:schemaLocation="http://diggsml.org/schemas/2.5.a https://raw.githubusercontent.com/DIGGSml/diggs-schema/2.5.a-dev/Complete
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://diggsml.org/schemas/2.5.a">
  - <documentInformation>
    - <DocumentInformation gml:id="DGS765F-DA6-2E5C-4987-1FA95">
        <creationDate>2018-12-11</creationDate>
      - <author>
        - <BusinessAssociate gml:id="BA_">
            <gml:name/>
          </BusinessAssociate>
        </author>
          <!--<author></author>-->
        <disclaimer>Not specified in gINT file</disclaimer>
      - <auditTrail>
        - <Remark>
            <content>Generated by Keynetix Diggs Feedback Tool</content>
            <remarkDateTime>2018-12-11</remarkDateTime>
          </Remark>
        </auditTrail>
      </DocumentInformation>
    </documentInformation>
  - <project>
    - <Project gml:id="Project_600162">
        <gml:identifier codeSpace="ODOT_CS">100056</gml:identifier>
        <gml:name>600162</gml:name>
          <!--NO roles defined for this file conversion<role> <Role> <rolePerformed>Client</rolePerfo
          </Role> </role>-->
      - <locality>
        - <Locality gml:id="pl_600162">
          - <address>
            - <Address gml:id="pla_600162">
                <number/>
                <streetAddress/>
                <city/>
                <state/>
                <province/>
                <county>PER</county>
                <country/>
                <postalCode/>
              </Address>
            </address>
```

Python variables

Variable explorer

Name	Type	Size	Value
extractedMeasurement	dict	2	{'B-001-0-15':{'pos':[...], 'name':[...], 'investigationTarget':[...], ...
extractedObservations	dict	2	{'B-001-0-15':{'lithologyClassificationType':[...], 'posList':[...], ' ...
extractedSampleInfo	dict	2	{'B-001-0-15':{'pos':[...], 'name':[...], 'projectRef':[...], 'samplin ...
extractedSamplingActivity	dict	2	{'B-001-0-15':{'pos':[...], 'name':[...], 'investigationTarget':[...], ...
extractedSamplingFeature	dict	2	{'B-001-0-15':{'Borehole':[...], 'rolePerformed':[...], 'BusinessAssoc ...
file_path	str	1	C:/Users/hwang12/Documents/Project_on_going/ODOT_subsurface_modeling/s ...

extractedMeasurement - Dictionary (2 elements)

Key	Type	Size	Value
B-001-0-15	dict	19	{'pos':['1.5; 3.5; 6; 8.5; 11; 12.5; 1.5; 6; 8.5; 1.5; 6; 8.5; 1.5; 3. ...
B-002-0-15	dict	19	{'pos':['1.5; 3.5; 6; 8.5; 11; 12.5; 18; 23; 6; 11; 6; 11; 1.5; 3.5; 6 ...

B-001-0-15 - Dictionary (19 elements)

Key	Type	Size	Value
DriveSet	list	3	['ds-1_DGS64EF-1C1-4F55-EE30-5D3E8', 'ds-2_DGS8F01-184E-1EB1-D575-2830 ...
DrivenPenetrationTest	list	1	['DGS931A-B66-2D49-2917-13D3A; DGS274E-110D-29D1-ECEA-2F6A4; DGS5561-D ...
LposList	list	1	['1.5 3; 3.5 5; 6 7.5; 8.5 10; 11 12.5; 12.5 22.5; 1.5; 6; 8.5; 1.5; 6 ...
PropertyParameters	list	1	['DGSA1E2-1103-4284-773F-2280E; DGS2408-B3D-22EF-5DC7-672D7; DGS81CF-1 ...
TestResult	list	1	['DGSAF8-283-3830-37C5-26608; DGS16F6-1121-A89-9A-2004E; DGS76A7-16A2- ...
blowCount	list	3	['6', '5', '4; 1; 4; 11; 53; None; None; None; None']
dataValues	list	1	['9; 3; 7; 18; 84; 0; 0 0; 31 20; 37 18; None; None; None; ,A-1-b,A-1- ...
index	list	3	['1', '2', '3; 3; 3; 3; 3; None; None; None; None']
investigationTarget	list	1	['Natural Ground; Natural Ground; Natural Ground; Natural Ground; Natu ...
name	list	1	['SPT; SPT; SPT; SPT; SPT; SPT; ATTERBERG; ATTERBERG; ATTERBERG; PARTI ...
penetration	list	3	['6', '6', '6; 6; 6; 6; 6; None; None; None; None']
penetrationTestType	list	1	['SPT; SPT; SPT; SPT; SPT; SPT; None; None; None']
pos	list	1	['1.5; 3.5; 6; 8.5; 11; 12.5; 1.5; 6; 8.5; 1.5; 6; 8.5; 1.5; 3.5; 6; 8 ...
projectRef	list	1	['#Project_600162; #Project_600162; #Project_600162; #Project_600162; ...
propertyClass	list	1	['n_value; n_value; n_value; n_value; n_value; n_value; plastic_limit; ...
propertyName	list	1	['N-Value; N-Value; N-Value; N-Value; N-Value; N-Value; Plastic Limit; ...
samplingFeatureRef	list	1	['#Borehole_B-001-0-15; #Borehole_B-001-0-15; #Borehole_B-001-0-15; #B ...
totalPenetration	list	1	['18; 18; 18; 18; 18; 120; None; None; None']
typeData	list	1	['long; long; long; long; long; long; integer; integer; integer; None; ...

Running time: 0.044s for this example file

Figure 4. The developed python interface (The .xml files are parsed and converted into customized python native data structures, which can be easily processed by other modules)

3. Case study and validation

3.1 Project CUY-480 – soil profile modeling and uncertainty quantification

3.1.1 Project CUY-480 introduction

The project is a deck replacement on the twin structures carrying IR 480 over the Cuyahoga River Valley. The bridge spans the CSX Railroad, the Cuyahoga River, Cleveland Metroparks Ohio and Erie Canal Reservation and Towpath Trail, West Canal Road, the Ohio Canal, and Canal Road. The location of the crossovers on the west end of the project is between the IR 480 Bridge over the IR 77 ramps and the SR 21 Bridge over IR 480. The location of the crossovers on the east end of the project is to the east of the Transportation Boulevard interchange. The plan view of the project CUY-480 is shown in Figure 5.

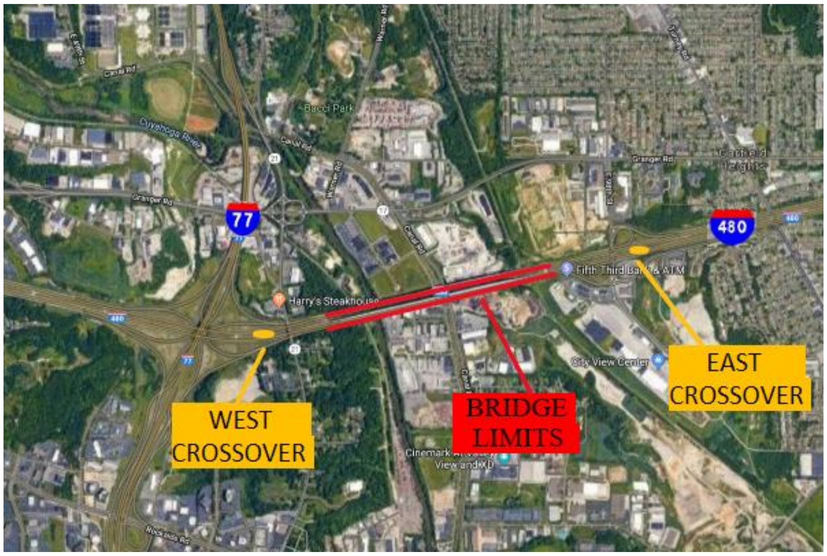


Figure 5. The plan view of the project CUY-480

Within the Bridge Limits region, DIGGS data schema was adopted to digitally record and store the geotechnical site investigation data. CUY-480 DIGGS data includes borehole data collected in 2013 and 2016, with 10 boreholes in 2013 and 15 in 2016.

3.1.2 Data conversion

The original DIGGS compliant .xml files are parsed and converted into python data structures with the developed DIGGS--python interface. In the base feature classes of DIGGS schema, *Sampling Feature*, *Measurement* and *Observation* include all required geotechnical information of boreholes. All the features are written in GML in the raw .xml file (text file).

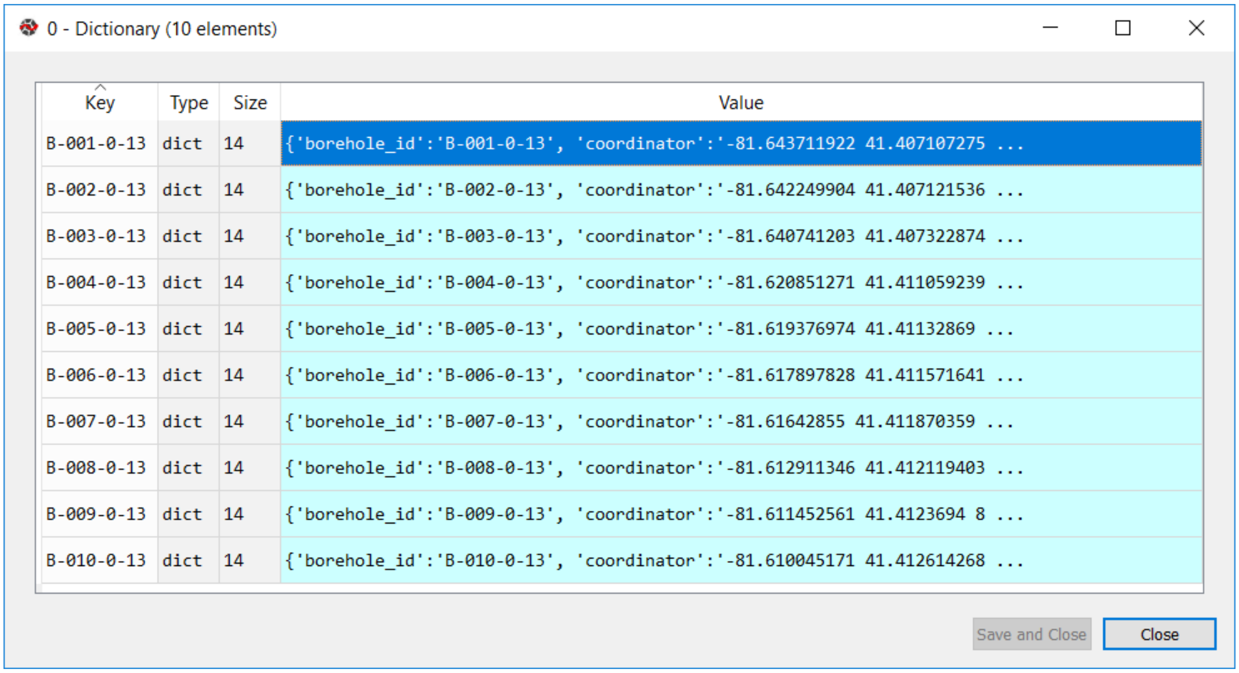
In the base feature class *Sampling Feature*, there are some subclasses: *Total-Measured Depth, Reference Point, Sampling Feature Property,* and *Borehole Construction Method*. From these subclasses, the borehole ID, borehole information of depth, latitude and longitude coordinators, and sampling method can be extracted. The subclass *Borehole Construction Method* contain two subclasses

Construction Method and *construction Equipment*. From the subclasses *Construction Method* and *Construction Equipment*, drilling method and drilling rig type can be identified respectively.

From the base feature class *Observation*, the layer thickness of each soil strata and the visual description of each soil layer can be extracted.

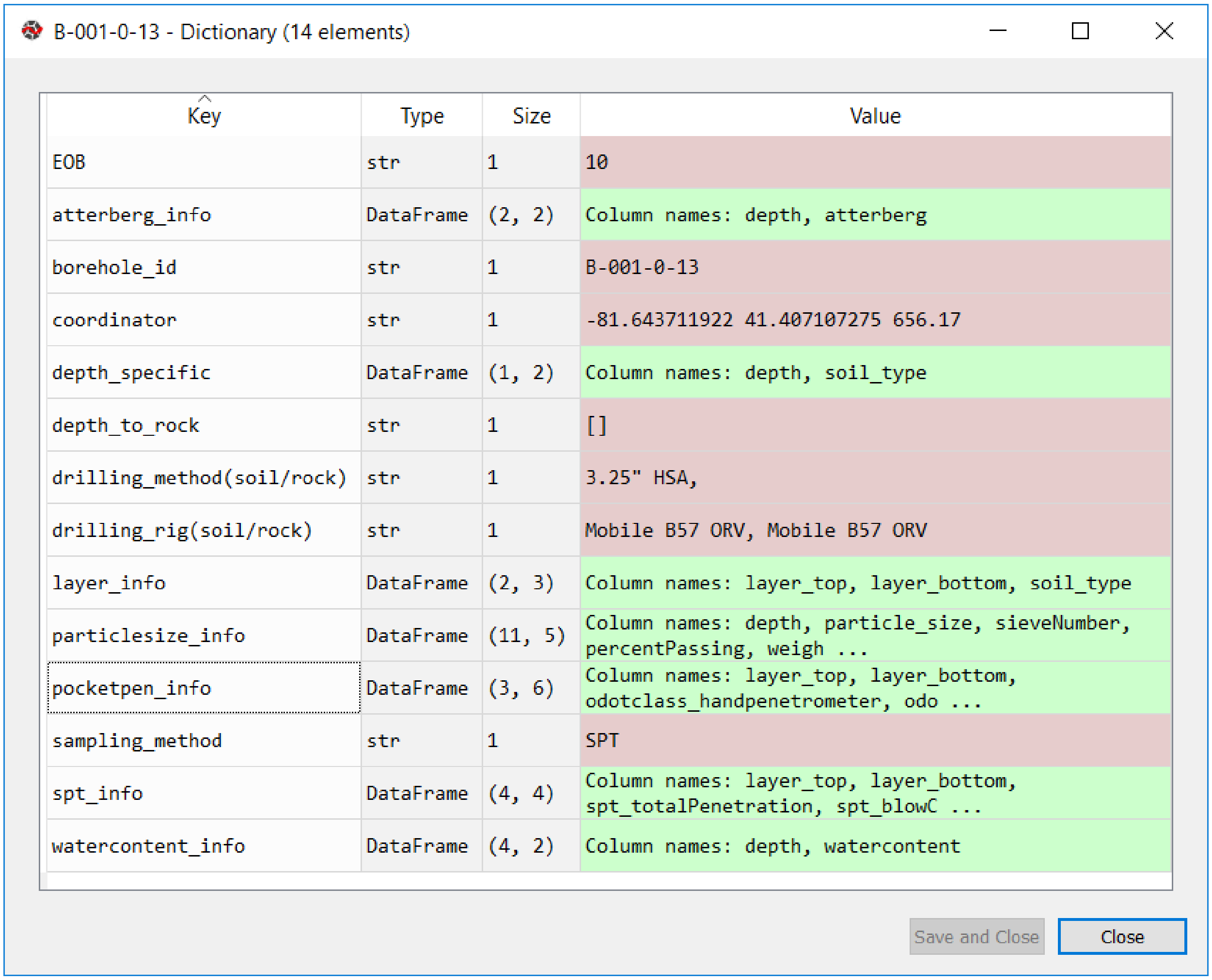
In the base feature class *Measurement*, there are three subclasses *Total Penetration*, *Blow Count* and *Position List*. For example, from the subclass *Total Penetration* and *Blow Count*, the soil penetration test blow count of each layer at each borehole can be extracted. Some physical properties also can be extracted from the subclasses *Particle Size* and *Sieve Number*.

By passing the raw .xml files into the interface, a list of borehole objects in native python *dictionary* data structure can be constructed as shown in Figure 6. The detailed information of each borehole is shown in Figure 7. Each item (defined by the “key”) in the *dictionary* is saved using a specific data type that will be further utilized in the modules of the developed software.



Key	Type	Size	Value
B-001-0-13	dict	14	{'borehole_id':'B-001-0-13', 'coordinator':'-81.643711922 41.407107275 ...
B-002-0-13	dict	14	{'borehole_id':'B-002-0-13', 'coordinator':'-81.642249904 41.407121536 ...
B-003-0-13	dict	14	{'borehole_id':'B-003-0-13', 'coordinator':'-81.640741203 41.407322874 ...
B-004-0-13	dict	14	{'borehole_id':'B-004-0-13', 'coordinator':'-81.620851271 41.411059239 ...
B-005-0-13	dict	14	{'borehole_id':'B-005-0-13', 'coordinator':'-81.619376974 41.41132869 ...
B-006-0-13	dict	14	{'borehole_id':'B-006-0-13', 'coordinator':'-81.617897828 41.411571641 ...
B-007-0-13	dict	14	{'borehole_id':'B-007-0-13', 'coordinator':'-81.61642855 41.411870359 ...
B-008-0-13	dict	14	{'borehole_id':'B-008-0-13', 'coordinator':'-81.612911346 41.412119403 ...
B-009-0-13	dict	14	{'borehole_id':'B-009-0-13', 'coordinator':'-81.611452561 41.4123694 8 ...
B-010-0-13	dict	14	{'borehole_id':'B-010-0-13', 'coordinator':'-81.610045171 41.412614268 ...

Figure 6. Converted borehole objects in native python data structure



Key	Type	Size	Value
EOB	str	1	10
atterberg_info	DataFrame	(2, 2)	Column names: depth, atterberg
borehole_id	str	1	B-001-0-13
coordinator	str	1	-81.643711922 41.407107275 656.17
depth_specific	DataFrame	(1, 2)	Column names: depth, soil_type
depth_to_rock	str	1	[]
drilling_method(soil/rock)	str	1	3.25" HSA,
drilling_rig(soil/rock)	str	1	Mobile B57 ORV, Mobile B57 ORV
layer_info	DataFrame	(2, 3)	Column names: layer_top, layer_bottom, soil_type
particlesize_info	DataFrame	(11, 5)	Column names: depth, particle_size, sieveNumber, percentPassing, weigh ...
pocketpen_info	DataFrame	(3, 6)	Column names: layer_top, layer_bottom, odotclass_handpenetrometer, odo ...
sampling_method	str	1	SPT
spt_info	DataFrame	(4, 4)	Column names: layer_top, layer_bottom, spt_totalPenetration, spt_blowC ...
watercontent_info	DataFrame	(4, 2)	Column names: depth, watercontent

Figure 7. Detailed extracted borehole information

Within the bridge limits as shown in Figure 5, 25 boreholes were logged, within which 8 (a-f, V-1 and V-2 in Figure. 8) were drilled and recorded using DIGGS data schema in 2013 and 2016, the rest 17 are archived borehole logs. The horizontal plan of all boreholes are shown in Figure 8. All borehole locations are projected onto the straight line connecting Borehole 1 and 17 and this straight line indicates the location of the targeted soil profile. Six DIGGS logs a-f and 17 archived logs are used for soil profile simulation, parameter estimation and uncertainty quantification while the remaining two boreholes (V-1,2) are used for validation. The latitude and longitude coordinates of each borehole location are converted into Universal Transverse Mercator (UTM) coordinates. In Figure 8, the blue open circles indicate the actual location of the boreholes logged in DIGGS format, the green open circles indicate the archived borehole locations, and the black solid circles are the projection of these boreholes on the reference line.

Profile along the line in Figure 8 and the soil stratification of each borehole is shown in Figure 9.

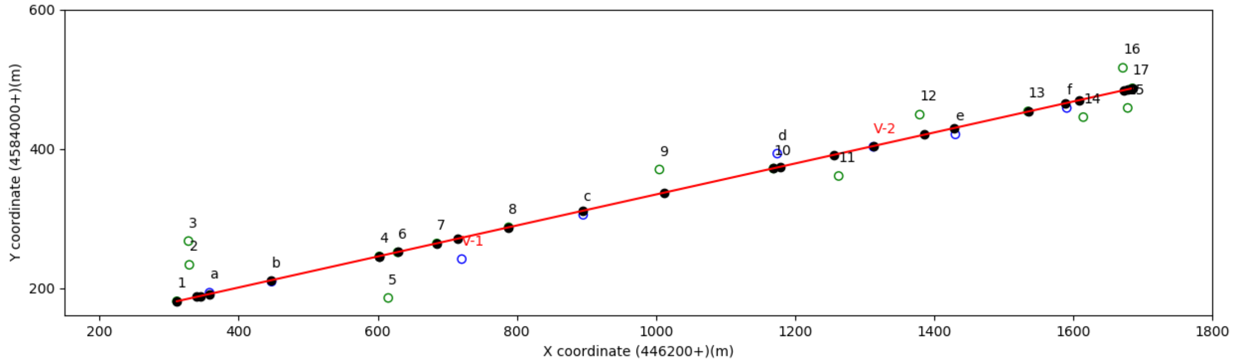


Figure 8. Plan view of all boreholes (x- y- coordinates are in Universal Transverse Mercator (UTM) coordinate system).

According to ODOT classification for geotechnical logging of soil and rock stratum, soil samples can be classified into and recorded as one of the following types: A-1-a, A-1-b, A-3, A-3a, A-2-4, A-2-5, A-2-6, A-2-7, A-4a, A-4b, A-5, A-6a, A-6b, A-7-5, A-7-6, A-8a, A-8b. For practical reasons and based on current ODOT practice, a more simplified classification of soil types is adopted in this report:

- A-1-a, A-1-b, A-3, A-3a, A-2-4, A-2-5, A-2-6, A-2-7 are classified as Non-Cohesive soil (NCO);
- A-6a, A-6b, A-7-5, A-7-6 are classified as Cohesive soil (COH);
- A-4a, A-4b are classified as Intermediate soil that between cohesive and non-cohesive soil (INT);
- A-8a, A-8b are classified as organic soil (ORG).

The vertical observed stratification at individual borehole location is shown in Figure. 9. It needs to be mentioned that the maximum elevation difference is 387 feet (118 meters) and the distance from borehole No.1 to No.17 is 4616 feet (1407 meters). The vertical profile is discretized into 100 (rows) by 201 (columns) pixels and thus each pixel indicates a 23 x 4 feet (7 x 1.18 meter) rectangular area in the vertical section.

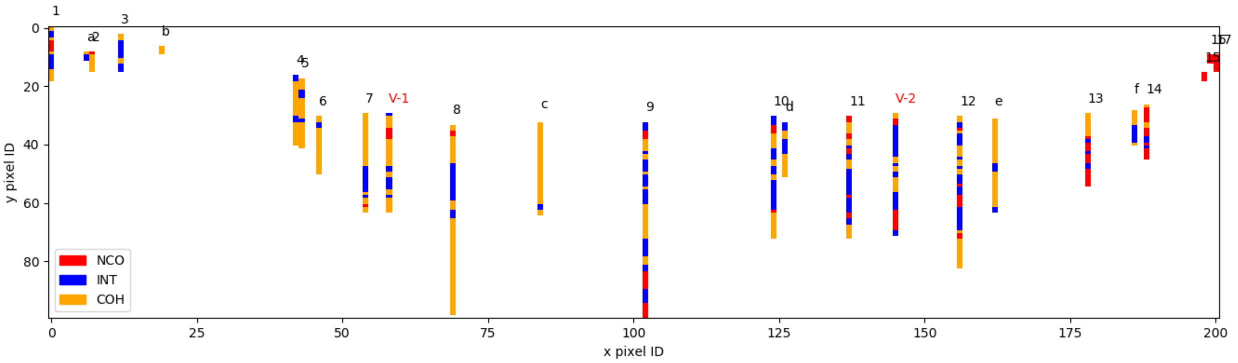


Figure 9. The observed vertical stratifications at all borehole locations.

3.1.3 Stochastic simulation

The known pixels that represent available borehole information (i.e., 1-17 and a-f) in Figure 9 are then transferred into the Module 3 of the developed software. 5000 stochastic simulations are performed to have robust estimates of the model parameters and converged soil profile realizations. Since the soil labels (i.e., the soil types) at unknown pixels are all inferred from the known pixels (i.e., the available boreholes) and prior information about the spatial anisotropy of the subsurface configuration, they are uncertain and have different preference of soil labels at different pixels. The uncertainty level of a specific pixel can be measured and reflected by the information entropy. A higher information entropy value signals larger uncertainty in the stratum assignment (i.e., the assigned soil/rock label) in the element. For example, the element that has the equal possibilities of being assigned with different strata could yield the highest information entropy value, which is the most uncertain scenario. It can be seen from Figure 10 that the stratigraphic uncertainty in the element close to the boreholes is relatively small and negligible; whereas, that in the element far away from the boreholes is large (i.e., the boundaries between adjacent strata become more uncertain). Thus, the borehole data exert a strong constraint on the stratum assignment in the nearby elements, this constraint, however, decreases with the distance measured from the nearest boreholes. The information entropy in Figure 10 might also imply that with the increase in the density of boreholes, the stratigraphic configuration at the site of concern could be more accurately characterized. It can be noticed that the information entropy map is heterogeneous and most of the pixels have entropy value greater than 0.5, which means the uncertainty level of the inferred pixels is high. This is due to the fact that only 23 sparse boreholes are known spanning across more than 1400 meters (only 8.4% pixels of the unmasked area in Figure 10 are known). More detailed discussion in this regard is provided below.

By analyzing and visualizing the spatial distribution of derived information entropy, we can quantitatively assess the difficulty of this soil profile delineating problem together with an estimation of the soil profile. In this investigation, two estimated soil profiles are reported; namely, maximum a posteriori (MAP) estimation, and the maximum log(target) realization. The former one shows the soil profile taking the uncertainty of fitted model parameters into consideration and the latter one is a specific realization using the sampled model parameter with the highest posterior probability density. The difference between the two soil profile estimations is two-fold: 1) the MAP estimation is a statistic of the entire random field while the maximum log(target) estimation is just a single realization; 2) the MAP estimation is derived from the marginal distribution of the soil label probability at each pixel while the log(target) estimation is only conditional on a specific parameter set.

The MAP estimation provides a map of the most probable soil type at a per-pixel basis. This map is a “smoothed” estimation of the entire random field by eliminating detailed local anisotropy effect introduced during the stochastic simulation process. On the other hand, the maximum log(target) estimation is a single realization using the most probable model parameters; hence, it can well preserve the local anisotropy yet lack consideration regarding the uncertainty of model parameters.

Actually, without uncertainty quantification, neither of them can represent the truth properly. The reason can be expressed from two different perspectives. First, given the information only at a limited number of borehole locations, the Markov random field parameters need to be inferred based on these known boreholes and some prior knowledge if available. The uncertainty introduced by the unknown parameters can strongly affect the stochastic simulation process as highly uncertain model parameters will prevent the simulation from converging (i.e., solving an ill-posed problem). In other words, there should exist a critical number of boreholes (corresponding to a given resolution of the discretized subsurface, and inferential algorithm) at a given site. No method can achieve a reasonable inference of the soil profile if the known borehole information is insufficient. Even the stochastic simulation can converge with very limited borehole information, a high overall information entropy level is expected, as frequently changing model parameters due to insufficient information during the simulation process can result in highly uncertain random field realizations. Under this condition, neither MAP or maximum log(target) estimation can represent the entire field. Second, because the distance between neighboring boreholes varies at a given site (e.g., the borehole spatial distribution in Figure 9) and the local complexity of the soil layer configuration may be changing across the entire site, the spatial distribution of information entropy is heterogeneous. The soil labels of pixels with low information entropy are more certain than those with high information entropy, and hence only a portion of the pixels in the MAP or maximum log(target) estimation having a low level of uncertainty can be considered as representative soil labels whereas other pixels are highly uncertain and, without additional boreholes, it is not meaningful to infer the true labels of these pixels in a statistical sense. This is why uncertainty quantification is of paramount importance in assessing the accuracy of a subsurface model.

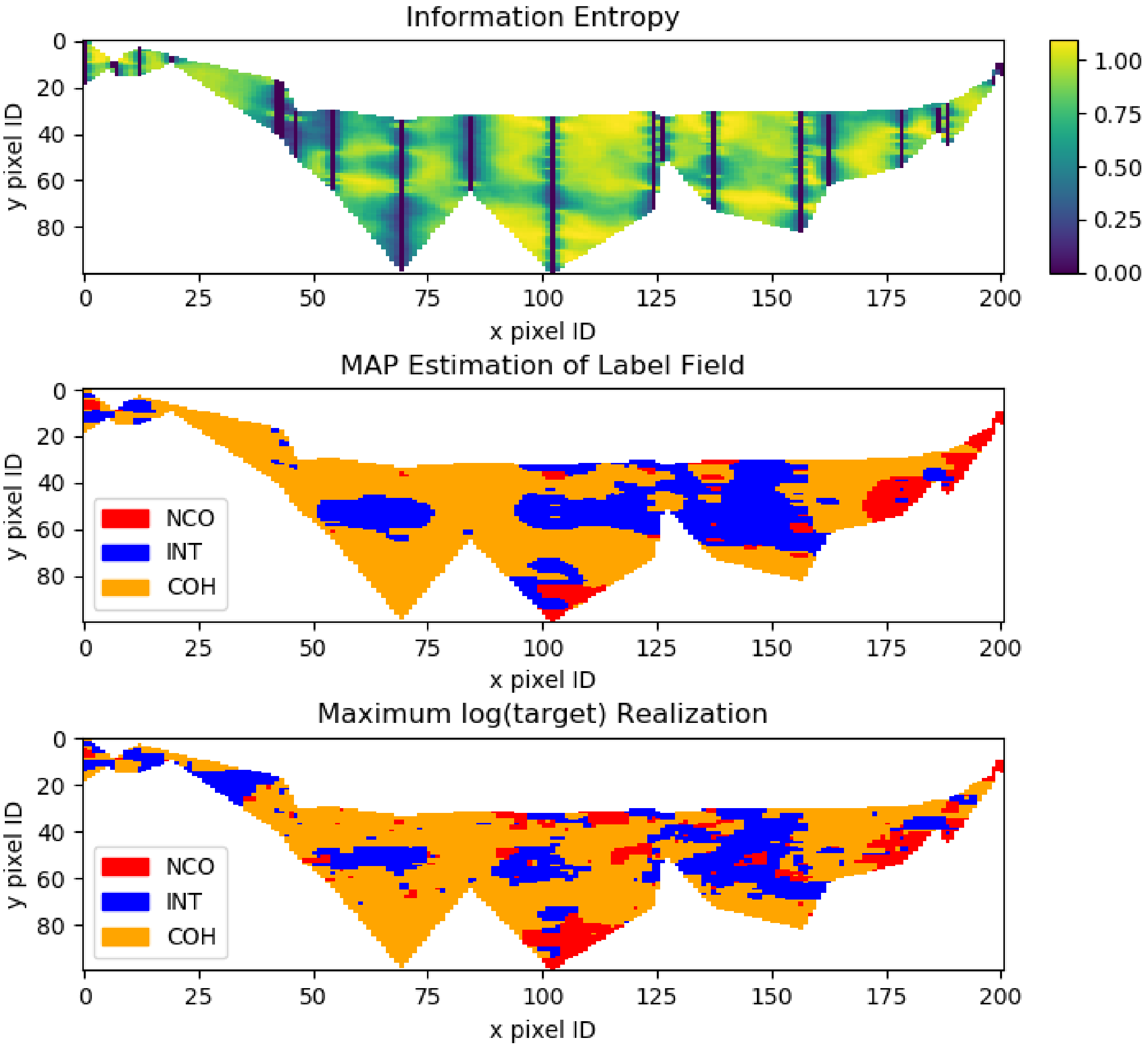


Figure 10. Simulation results: soil profile estimation and the associated uncertainty quantification using information entropy.

3.1.4 Validation

For a detailed validation, the MAP estimation at the pixels of the validation boreholes is extracted and compared with the actual observations logged in the DIGGS file, the comparison results are shown in Figure 11. V-2 prediction has a higher overall uncertainty level than that of V-1 and hence the prediction at V-1 has higher accuracy. The predictions are generally smooth and cannot detect detailed thin layers, however, the information entropy may provide some indication of the local complexity. By comparing the MAP estimation with the truth, 72.7% of V-1 is correctly inferred, and 53.7% of V-2 is correctly inferred. The much lower inferential accuracy at V-2 can be expected as can be noticed from the ground truth and the overall inferred profile from Figure 10, the dominant soil type at V-2 is “INT”. This type of soil is in the “fuzzy zone” between the “NCO” and “COH”, besides, the “hard” classification of soil samples at borehole locations during the labeling process did not provide enough information regarding the uncertainty of possible soil types. This issue is even more pronounced when it comes to differentiating the “COH” from “INT” based on laboratory testing results (e.g., the subtle difference between a Plasticity Index of 10 vs. 11 would result in a “INT” soil vs. a “COH” soil). The other reason could be the local complexity. Suppose V-2 is not known in advance, the neighboring Boreholes 11 and 12 have frequent changes of soil types along depth and can result in high information entropy in the region between them (Figure 10). In this regard, the inferred MAP estimate is not reliable anymore as no dominant soil type can be determined and multiple soil types may have comparable likelihood. This point also can be reflected by the noticeable differences between the MAP estimate and the maximum log(target) realization in this region.

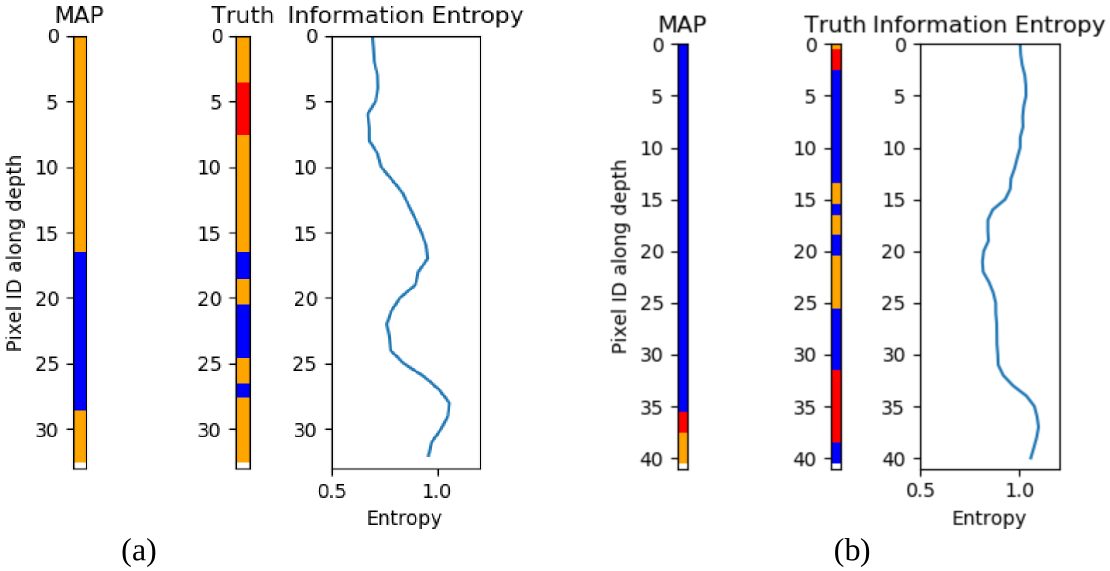


Figure 11. Validation results using two validation boreholes: (a) V-1; (b) V-2.

For the sake of acquiring more objective validation results, leaving-one-out cross validation is adopted here. To be specific, each and every borehole is considered as the validation borehole in turn and the rest ones are used for simulation. Then the simulated soil stratification at the borehole location is compared with the true stratification at the validation borehole. The validation result is shown in Table 1.

As shown in Table 1, the accuracy rate of boreholes 15, 16, 17 and b are 100%, the accuracy rate of borehole 2 is 0%. However, it needs to be noticed that these boreholes are very short and hence less representative. When checking other boreholes, the accuracy rate of boreholes 4, 5, 6, 7, V-1 and e are over 60%, the accuracy rate of boreholes 11, V-2, 13, 14, 15 and c are poor and less than 40%. The reason for the good inferencing results is that the validation borehole is close to the adjacent boreholes, while the validation boreholes in the latter are relatively far from the known borehole information and hence subject to softer spatial constraints. The accuracy rate of boreholes d and f are low, though the adjacent boreholes on both sides of d or f are close. The local complexity is higher at these locations, and hence without additional exploration, it is difficult to have accurate inference.

Table 1. Validation result of each borehole

Validation Borehole ID	Minimum Entropy	Maximum Entropy	Mean Entropy	Borehole Length(m)	Accuracy Rate
1	1.07	1.1	1.08	17.29	0.37
2	0.76	0.92	0.85	6.37	0.0
3	1.01	1.09	1.05	11.83	0.31
4	1.02	1.1	1.06	21.84	0.75
5	0.7	0.97	0.81	21.84	0.83
6	1.0	1.06	1.03	18.2	0.9
7	1.03	1.1	1.08	30.94	0.76
V-1	0.58	1.01	0.85	30.94	0.68
8	1.01	1.1	1.06	59.15	0.46
9	1.04	1.1	1.08	60.97	0.4
10	0.74	1.09	0.95	38.22	0.36
11	1.07	1.1	1.09	38.22	0.31
V-2	1.07	1.1	1.09	38.22	0.43
12	0.87	1.06	1.01	47.32	0.4
13	1.06	1.1	1.08	22.75	0.36
14	0.91	1.09	1.04	17.29	0.32
15	1.03	1.05	1.04	2.73	1.0
16	0.73	0.76	0.74	2.73	1.0
17	0.56	0.9	0.73	5.46	1.0
a	0.54	0.98	0.69	2.73	0.33
b	1.09	1.1	1.09	2.73	1.0
c	1.04	1.1	1.08	29.12	0.25
d	1.02	1.1	1.06	17.29	0.26
e	1.05	1.09	1.07	28.21	0.81
f	0.93	1.08	1.02	10.01	0.27

The accuracy also be partially reflected by the information entropy. The minimum information entropy of boreholes 2, 5, V-1, 10, 16, 17 and a are less than 0.8, and the mean information entropy of boreholes 16, 17 and a are also less than 0.8. It indicates that the uncertainty of their simulation results is relatively low although they are very short and less representative.

3.2 Project CUY-IR-490/SR10-02.09/19.28 – soil profile modeling and uncertainty quantification

3.2.1 Project introduction

The second case study is also a site exploration project for roadway bridge structure foundation design. Only limited information is available for this case study including boring locations, boring logs, and lab data of soil samples. For the subsurface modeling and uncertainty quantification, only boring plan, boring logs are used in this demonstration. The site exploration project was conducted in 2014 and 16 borehole logs were recorded, within which 14 boreholes are close to the roadway alignment and selected for further soil profile delineation and validation. The plan view with UTM coordinates is shown in Figure 12. All borehole locations are project onto the red line indicating the roadway alignment. The observed soil types at borehole locations are simplified into four categories (i.e., NCO, COH, INT, ORG) as defined in the previous example. The vertical observed stratification at individual borehole location is shown in Figure 13.

The maximum elevation difference is 177 feet (54 meters) and the distance from borehole No.2 to No.12 is 617 feet (188 meters). The vertical profile is discretized into 54 (rows) by 101 (columns) pixels and thus each pixel indicates a 6 x 3.3 feet (1.86 x 1 meter) rectangular area in the vertical section.

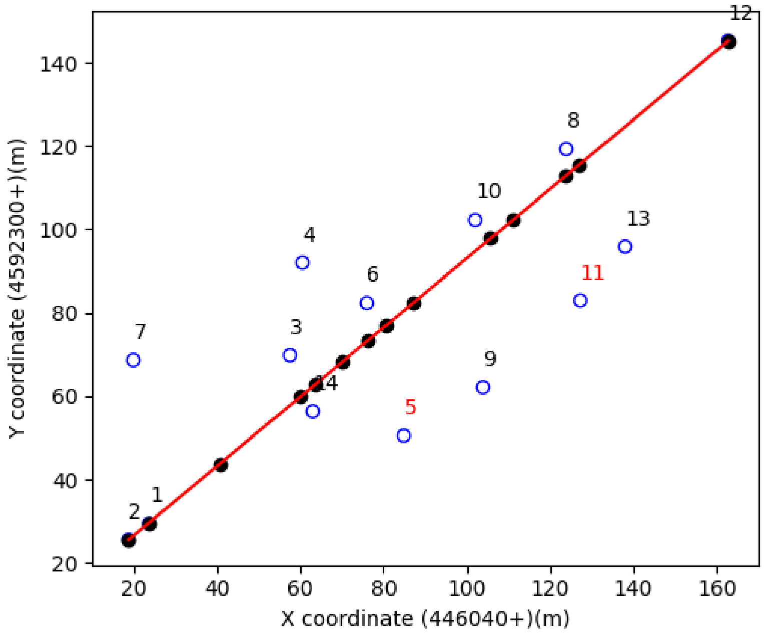


Figure 12. Plan view of all boreholes (x- y- coordinates are in Universal Transverse Mercator (UTM) coordinate system).

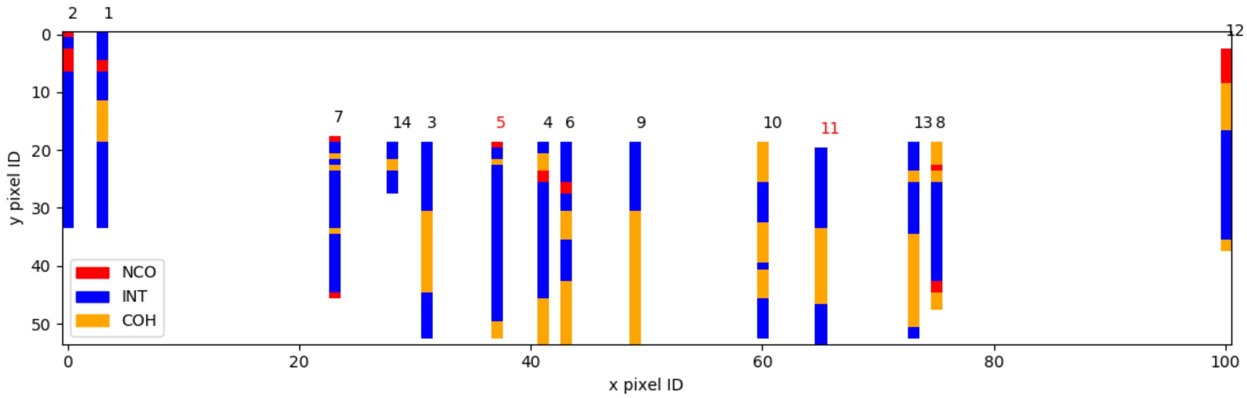


Figure 13. The observed vertical stratifications at all borehole locations.

3.2.2 Stochastic simulation

In this example, the horizontal spatial range of the simulated soil profile is much smaller than the previous example (617 feet (188 meters) in this example versus 4616 feet (1407 meters) in the previous example). In order to have a higher spatial resolution, as mentioned above, the actual width of each pixel is 6 feet (1.86 meter) versus 23 feet (7 meters) in the previous example.

Two boreholes (No. 5 and No. 11) are used for validation and the rest for stochastic simulation of the soil profile. The same experimental procedure of the previous example was performed. The uncertainty quantification and simulation results are shown in Figure 14. As can be noticed, the overall uncertainty level is lower than previous example and most pixels have the entropy level lower than 0.8. Some artifacts (sharp and straight horizontal or vertical boundaries between different soil types) can be noticed in both MAP estimation of maximum log(target) realization. There are two possible reasons: 1) the stationary assumption may not be fully satisfied as real world soil spatial distribution should be non-stationary and heterogeneous; 2) subject to the number of sparse known boreholes, it is difficult to increase the resolution yet still have a well converged soil profile image (the low resolution can pronounce the effect resulted from the stationary assumption) Generally speaking, both of the two estimations are reasonable.

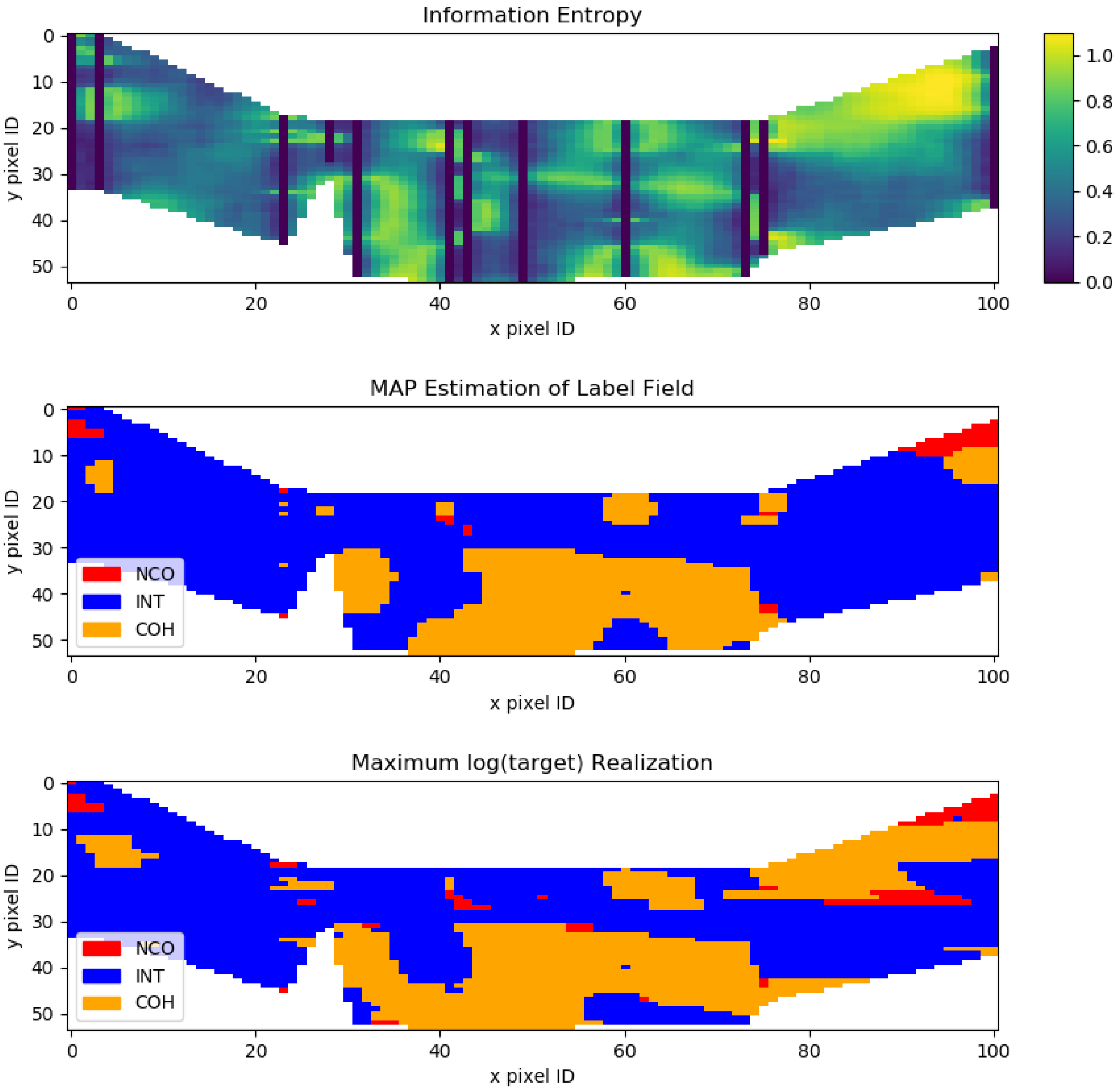
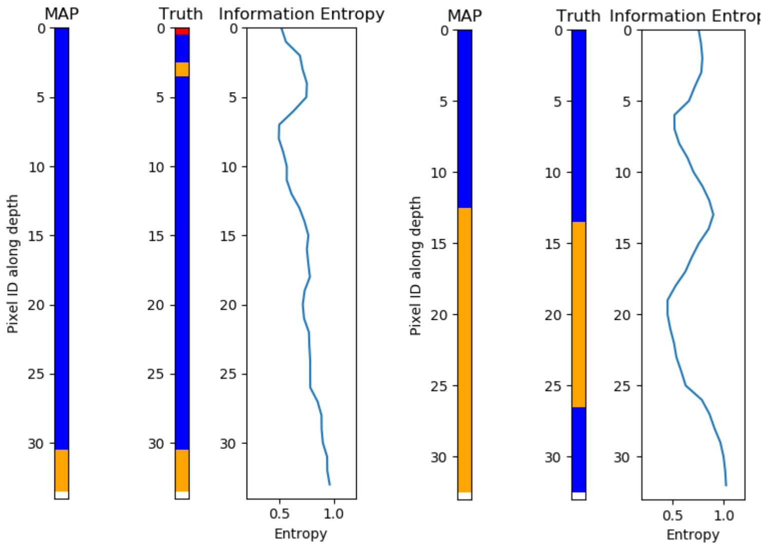


Figure 14. Simulation results: soil profile estimation and the associated uncertainty quantification using information entropy.

3.2.3 Validation

The two validation boreholes are then investigated by comparing the MAP estimation with observed borehole log data. The comparison results are shown in Figure 15. The overall accuracy of borehole No. 5 is 94.1% whereas the accuracy of borehole No. 11 is 78.8%. Similar as previous example, the general soil stratification can be inferred yet some details are missing since the local complexity cannot be well inferred only based on nearby boreholes.



(a) (b)

Figure 15. Validation results using two validation boreholes: (a) N0. 5; (b) No. 11.

A leave-one-out cross validation has been conducted for this project as well and the result is shown in Table 2.

Table 2. Validation results for each borehole of CUY-IR-490/SR10-02.09/19.28

Validation Borehole ID	Minimum Entropy	Maximum Entropy	Mean Entropy	Borehole Length(m)	Accuracy Rate
1	0.55	1.01	0.81	30.94	0.65
2	0.81	1.04	0.93	30.94	0.74
3	0.47	0.76	0.61	25.48	0.82
4	0.27	0.54	0.39	8.19	0.78
5	0.34	1.01	0.54	30.94	0.59
6	0.31	0.95	0.66	30.94	0.76
7	0.37	0.92	0.63	31.85	0.8
8	0.17	0.96	0.5	31.85	0.71
9	0.68	1.08	0.87	31.85	0.94
10	0.44	1.05	0.8	30.94	0.53
11	0.72	1.01	0.85	30.94	0.82
12	0.28	1.1	0.61	30.94	0.53
13	0.45	0.92	0.59	26.39	0.41
14	0.85	1.06	1.01	31.85	0.23

As shown in Table 2, the accuracy rate is generally high (9 out of 14 are higher than 0.6). Only boreholes 13 and 14 have relatively lower accuracy rate which are less than 0.5. Except borehole No.4, all other boreholes have a depth more than 82 feet (25 meters) and hence the accuracy rate is representative.

It can be noticed that the distance between neighboring boreholes has a great influence on the accuracy rate of inferred soil profile, which strongly agrees with our intuition. In addition, the resolution of the discretized subsurface section also matters, it needs to be compatible with the number of available boreholes so that certain amount of pixels have pre-assigned soil labels. This point is extremely important as mentioned above, only a sufficient number of known pixels can result in a converged simulation result. If this requirement cannot be satisfied, additional site investigation is required in order to have a reasonable soil profile inferencing result. In the current project, there are 100 × 54 pixels across the physical domain with 14 known boreholes. Yet there are much more pixels (200 × 100 pixels) in the previous case with 25 known boreholes. Both of them can result in converged results, however, the density of known information for project CUY-480 is much lower than that of current project. Thus, the overall level of information entropy of the current project is lower than that of project CUY-480.

3.3 Project CUY-480 - CPT joint interpretation

3.3.1 CPT dataset introduction

The CPT dataset of interest in this study consists of nine CPT sounding records that were collected in the years of 2015 and 2016 for roadway bridge structure foundation design in Project CUY-480. As shown in Figure 16, the available CPTs can be divided into two groups based on their locations. The first group consists of four CPTs drilled closely on the west embankment of the Cuyahoga River, with horizontal clearances vary from 0.59 m to 3.89 m; the second group consists of five CPTs drilled on the north embankment of the Erie Canal, with horizontal clearances vary from 1.08 m to 3.09 m. The drilling depths of these CPTs vary from 10.5 m to 39.6 m. All soundings were completed within either pre-cored holes or started with the use of a dummy cone to make a pilot hole due to the presence of urban fill material within the near-surface; the pre-drilling depths vary from 1.8 m to 3.0 m. An identical vertical sounding interval of 0.02 m applies for all the CPTs, which determines the size of the discretization lattice along the vertical direction in the proposed CPT joint interpretation approach.

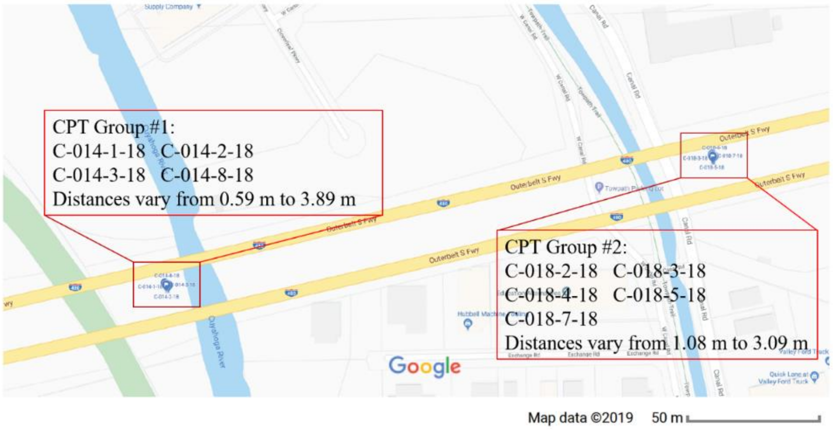


Figure 16. The plan view of the CPT locations in project CUY-480

3.3.2 Joint interpretation

As the horizontal spacing between these two CPT groups is more than 656 feet (200 meters), the horizontal correlations of the CPT soundings from the two groups can be neglected. Thus, two joint interpretation cases are performed for the two CPT groups. The ground level at the multiple CPT locations in each group is the same; therefore no further adjustment is applied for the depths of the CPT sounding records. The sounding points collected from the top ten-foot (three-meter) soil segments in the CPT records are discarded to address the various pre-drilling depths of the different CPT locations and to eliminate the potential impact of the pavement or backfilled materials. The pairwise sounding points, i.e., $\log_{10}F_r$ and $\log_{10}Q_{tn}$, are computed using the remaining raw CPT soundings. The computed sounding points from CPT Group #1 and Group #2 are combined and plotted as scatter diagrams in the conventional Robertson SBT_n chart, which is also referred to as the feature space in the developed approach, as shown in Figure 17(a) and (b), respectively. Cluster analyses can then be performed for each interpretation case using the developed joint interpretation algorithm, which considers not only the statistical patterns of the sound points in the feature space (e.g., the Robertson SBT_n chart), but also the spatial correlations among the sounding points in the 3-D physical subsurface space. As shown in Figure 17, the pairwise CPT sounding points in each interpretation case are categorized into seven and eight clusters for CPT Group #1 and Group #2, respectively. The estimated model parameters and the soil classes of each cluster are listed in Table 2. It is worth to mention that the total number of clusters is automatically estimated using the interpretation algorithm, instead of pre-defined, for each interpretation case. Based on the clustering result, the soil type of each CPT sounding point can be easily determined according to its associated cluster. Thus, the estimated stratification results for all the CPT soundings can be readily extracted simultaneously.

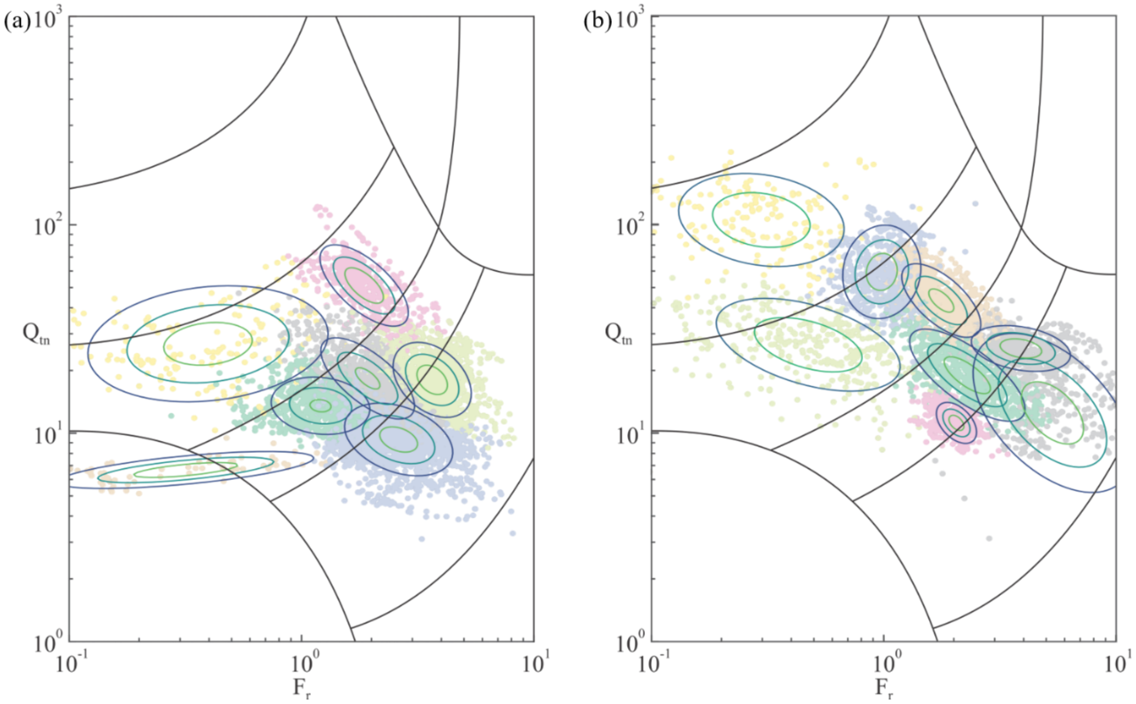


Figure 17. Clustering results for Project CUY-480 CPT dataset: (a) CPT Group #1; (a) CPT Group #2.

Table 2(a). Estimated model parameters for CPT Group #1 in Project CUY-480

Cluster #	SBT$_n$ #	Parameter	$\log_{10}F_r$	$\log_{10}Q_{tn}$
1	4	Mean	0.3036	1.2525
		Std	0.0039	0.0042
2	5	Mean	0.2758	1.6962
		Std	0.0036	0.0044
3	5	Mean	-0.3960	1.3292
		Std	0.0280	0.0132
4	4	Mean	0.1234	1.1391
		Std	0.0061	0.0035
5	1	Mean	-0.4829	0.8248
		Std	0.0409	0.0069
6	3	Mean	0.4148	0.9702
		Std	0.0057	0.0145
7	3	Mean	0.5632	1.2556
		Std	0.0041	0.0082

Table 2(b). Estimated model parameters for CPT Group #2 in Project CUY-480

Cluster #	SBT$_n$ #	Parameter	$\log_{10}F_r$	$\log_{10}Q_{tn}$
1	4	Mean	0.3599	1.2767
		Std	0.0018	0.0024
2	5	Mean	-0.0085	1.7721
		Std	0.0072	0.0061
3	4	Mean	0.3131	1.0480
		Std	0.0007	0.0020
4	5	Mean	-0.3252	1.4226
		Std	0.0166	0.0082
5	6	Mean	-0.5271	2.0207
		Std	0.0236	0.0126
6	5	Mean	0.2478	1.6346
		Std	0.0076	0.0037
7	3	Mean	0.7266	1.0959
		Std	0.0174	0.0427
8	4	Mean	0.5905	1.4048
		Std	0.0030	0.0053

3.3.3 Validation and discussion

The stratification interpretation results obtained using the developed joint interpretation algorithm for CPT Group #1 and Group #2 are presented in Figure 18 and Figure 19, respectively. For each of the CPT soundings, the corresponding interpretation results obtained using the conventional SBT$_n$ chart is also provided for comparison purpose. It can be noted from Figure 18 and 19 that the interpretation results obtained using the developed algorithm in general agree with the ones obtained using the conventional SBT_n chart, as the major soil layers identified using both methods locate in similar depths. However, a major difference is that the results obtained using the SBT_n chart exhibit excessively frequent changes of soil types along the CPT penetration paths, while the soil layers identified by the developed algorithm are much more continuous. In engineering practice, the extreme thin soil layers identified using the SBT_n chart are generally unrealistic, and thereby requires further manual inspection and simplification. However, interpretation of a site-scale CPT dataset consisting of numerous CPT records can be a cumbersome and time-consuming task for practicing engineers; and still, the revision process is subjective as it relies highly on the individual experiences and so-called engineering judgement. In contrast, the developed joint CPT interpretation algorithm can exploit the hidden spatial correlation among CPT sounding points and provide more accurate and reasonable interpretation results by automatically eliminating the unrealistic thin soil layers. More importantly, based on a horizontal comparison of the obtained stratification results of the CPT soundings in each group, it can be found that the results obtained using the developed joint interpretation algorithm is much more consistent along the horizontal direction. To be specific, it can be noted that the detected soil layer boundaries at different CPT spots locate at almost the same depths. Such strong consistency is a solid demonstration of the accuracy of the developed interpretation algorithm, since the CPTs in each group are drilled closely to each other, and a strong horizontal correlation of the soil stratification can be expected.

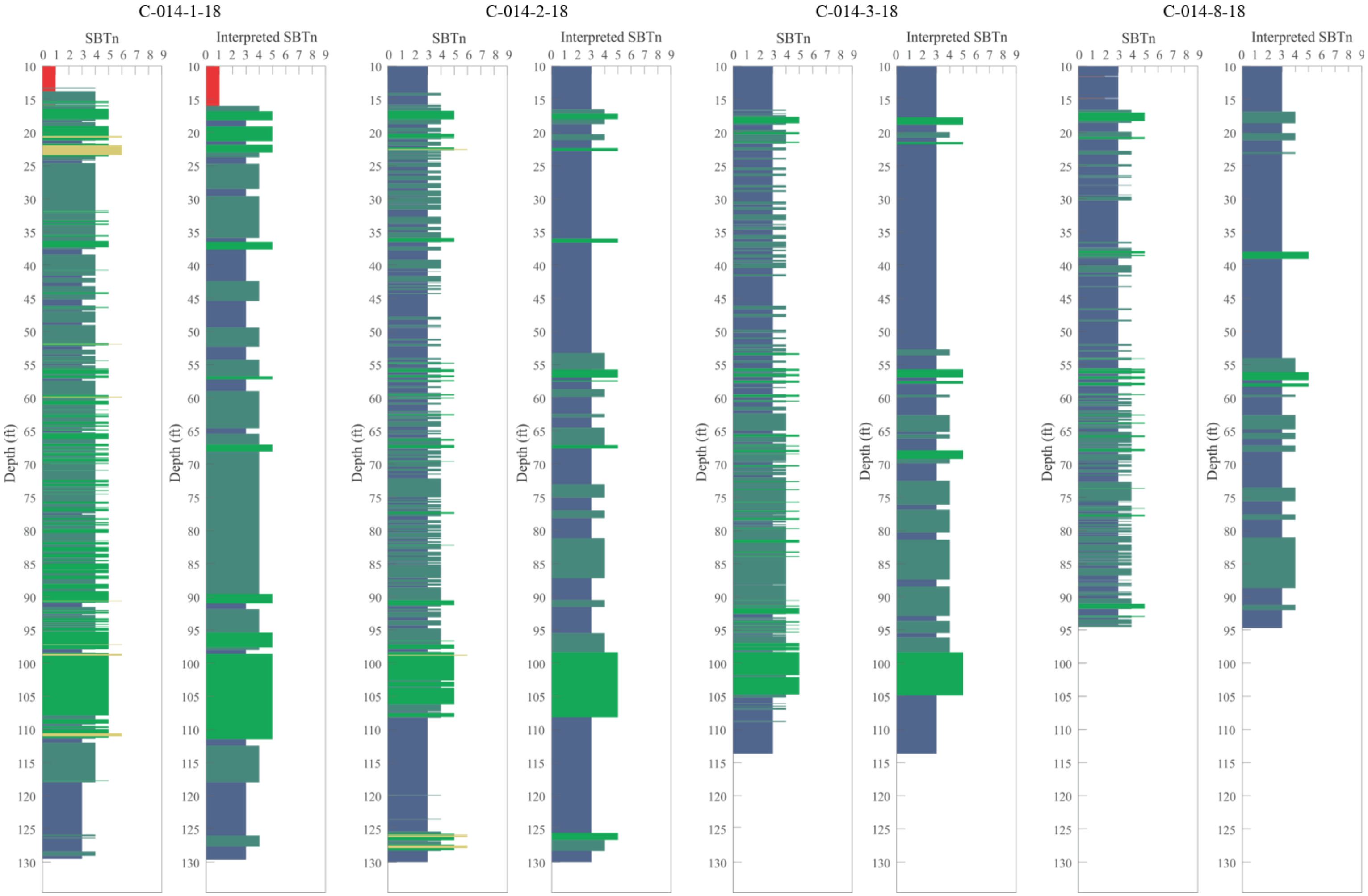


Figure 18. Stratification interpretation results for CPT Group #1 in the Project CUY-480.

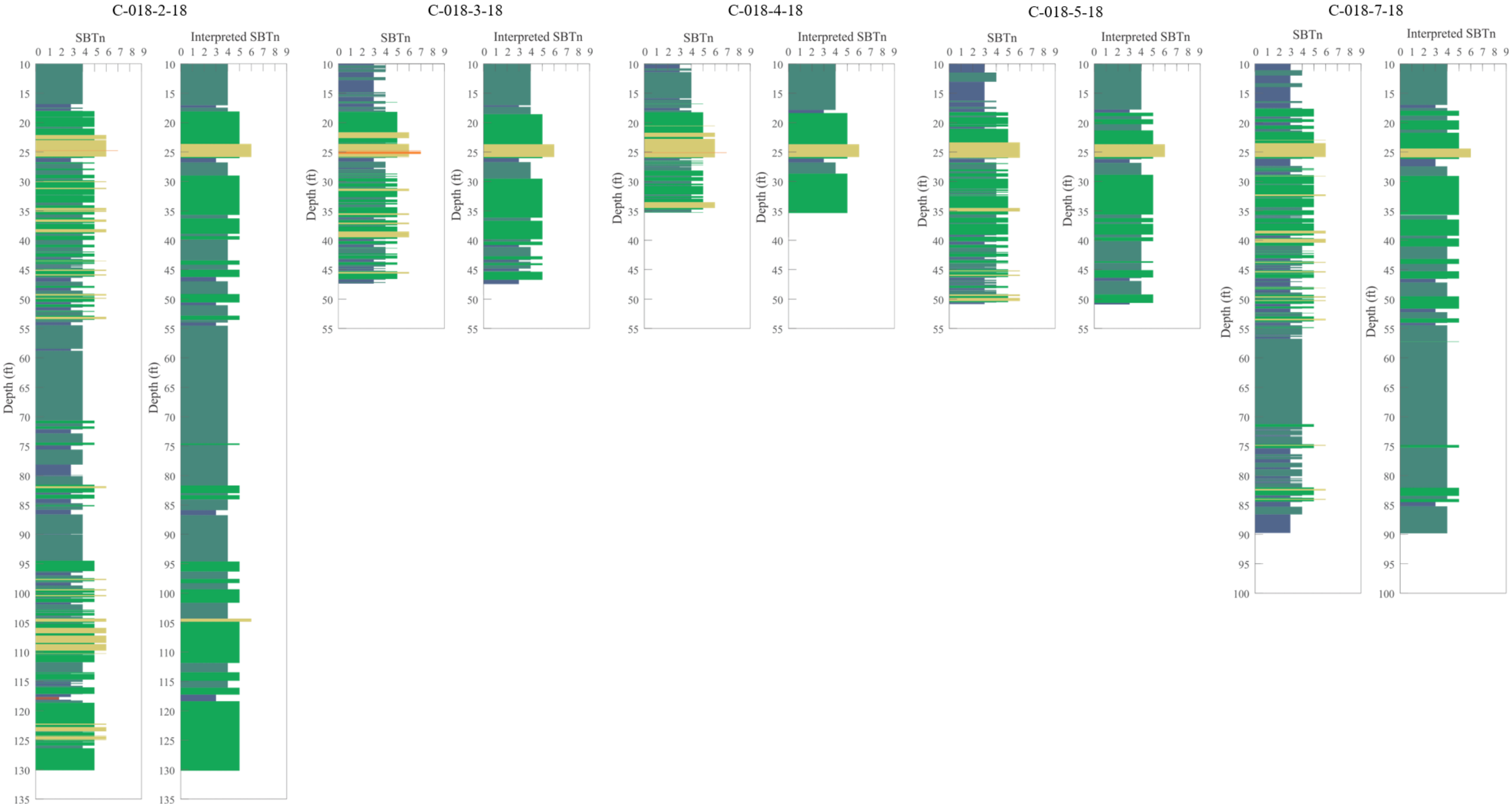


Figure 19. Stratification interpretation results for CPT Group #2 in the Project CUY-480.

3.4 Christchurch CBD Area - CPT joint interpretation

3.4.1 CPT dataset introduction

In this case study, a CPT dataset collected at a project site in Christchurch, New Zealand, is studied. The analyzed site is located within the central business district of the city of Christchurch, which is largely built upon a Late Quaternary substrate of gravel, sand, silt, and swamp deposits. The plan view of this project is shown in Figure 20. The studied dataset consists of 44 CPT soundings and three borehole logs that are collected from a 240 m × 240 m square region. The detailed CPT and borehole data are available through the Canterbury Geotechnical Database. The drilling depth of these CPTs vary from 15.13 m to 22.72 m; pre-drilling depths of 0.8 m to 1.5 m are conducted for specific CPT locations. An identical vertical sounding interval of 0.01 m applies for all the CPTs, which determines the size of the discretization lattice along the vertical direction in the proposed CPT joint interpretation approach. Three available boreholes (see Borehole #1- #3 in Figure 20) are located adjacent to CPT #6, #7, and #12, with horizontal distances of 2.6 m, 4.5 m, and 3.2 m, respectively. Thus, in this study, the drilling logs from these three boreholes are considered as the real stratification configuration at the corresponding CPT locations. Note that the available borehole information is not used as input for the CPT interpretation algorithms, but only serves as a validation set to evaluate the accuracy of the CPT interpretation results.

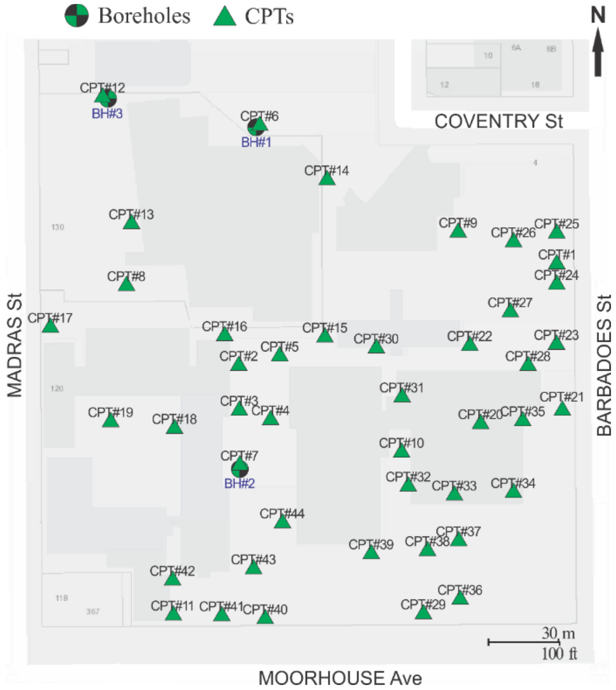


Figure 20. The plan view of the studied CBD Area in Christchurch, New Zealand

3.4.2 Joint interpretation

As the changes of the ground level at the CPT and borehole location are not significant (vary from 4.2 m to 4.6 m), it is assumed that the ground level at the multiple CPT location is the same, thereby no further adjustment is applied for the drilling depths of the CPT records and borehole logs. The sounding points collected from the top 1.5 m soil segments in the CPT records are discarded to address the various pre-drilling depths of the different CPT locations and to eliminate the potential impact of the pavement or backfilled materials. The pairwise sounding points, i.e., $\log_{10}F_r$ and $\log_{10}Q_{tn}$, are computed using the remaining raw CPT soundings. The computed sounding points from all the 44 CPT sounding records are plotted as a scatter diagram (see Figure 21) in the conventional Robertson SBT_n chart. A cluster analysis can then be performed using the developed joint interpretation algorithm. As shown in Figure 21, a total number of seven clusters are identified in the scatter diagram of the pairwise sounding points; the detailed estimated model parameters and soil classes of each cluster are listed in Table 3. Based on the clustering result, the soil type of each CPT sounding point can be easily determined according to its associated cluster. Therefore, the estimated stratification results for all the CPT soundings can be readily extracted simultaneously. In this study, we focus on two interpretation cases. In case #1, the accuracy of the two approaches for interpreting multiple CPT records will be evaluated by validating the respective stratification interpretation results at the locations of CPT #6, #7, and #12, using the corresponding borehole logs as the ground truths. In case #2, the interpretation consistency of the two approaches will be further inspected by comparing the respective stratification interpretation results at the locations of CPT #1 and CPT #24, which is the pair of CPTs with the shortest horizontal distance of 9.2 m at the studied site.

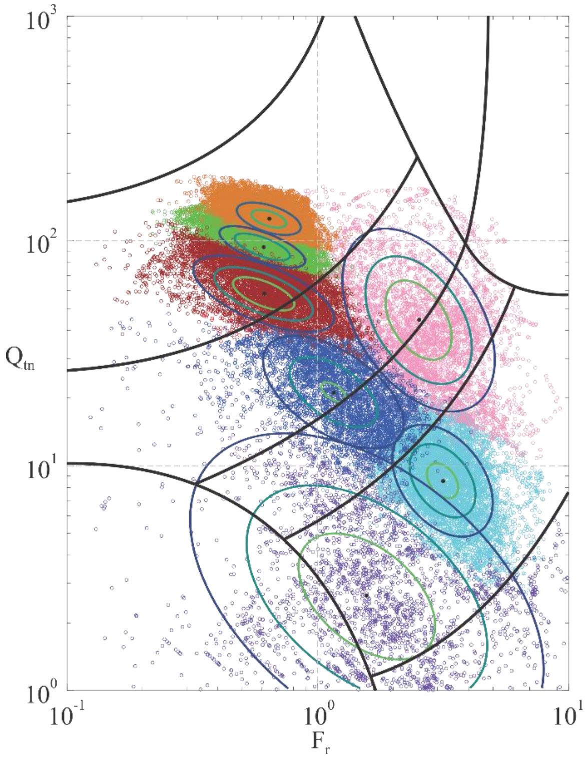


Figure 21. Clustering results for Christchurch CBD area CPT dataset.

Table 3. Estimated Model parameters using the joint interpretation approach

Cluster #	SBT_n #	Parameter	$\log_{10}F_r$	$\log_{10}Q_{tn}$
1	6	Mean	-0.1822	2.1092
		Std	0.0892	0.0520
2	6	Mean	-0.2140	1.9763
		Std	0.1104	0.0601
3	6	Mean	-0.2187	1.8905
		Std	0.1635	0.1000
4	5	Mean	0.3717	1.6323
		Std	0.1858	0.2417
5	5	Mean	0.0469	1.3426
		Std	0.2072	0.1626
6	3	Mean	0.4977	0.9455
		Std	0.1357	0.1545
7	3	Mean	0.1992	0.4333
		Std	0.4417	0.4482

3.4.3 Validation and discussion

We first examine the stratification interpretation results at the locations of CPT #6, #7, and #12, using the available boring logs as references. The interpretation results and the corresponding borehole logs are presented in Figure 22-24 to enable a close comparison. In these figures, the processed pairwise CPT sounding points (i.e., $\log_{10}F_r$ and $\log_{10}Q_{tn}$) are plotted in the first column; the stratification interpretation results obtained using the conventional SBT_n chart are shown in the second column for comparison purposes; the stratification interpretation results from the joint analysis approach are visualized in the third column; the stratification profile obtained from the available borehole logs are presented in the fourth column. It can be noted from Figure 22-24 that the developed joint interpretation algorithm can automatically eliminate the extraneous thin layers obtained using the conventional SBT_n chart. Meanwhile, focusing on the local differences between the stratification inferences obtained using the two approaches, it can be found that stratification results obtained using the developed algorithm is more consistent with the boring logs; thus, the joint interpretation approach can be considered as more accurate.

Then, we examine the stratification interpretation results at the locations of CPT #1 and CPT #24, which have the closest horizontal clearance among all the CPT pairs in the studied site. Figure 25 shows the interpretation results for CPT #1 and CPT #24, in which the processed pairwise sounding points of CPT #1and #24 are plotted in the first two columns. The stratification interpretation results obtained using the conventional SBT_n chart are shown in the third and the fourth columns for comparison purposes. The stratification interpretation results of these two CPTs obtained using the joint analysis approach are visualized in the last two columns. It can be noted from Figure 25 that the stratification results obtained using the joint interpretation approaches at these two CPT drilling locations are quite consistent with each other. The sounding points collected from two testing locations are labeled with the same set of soil classes, and the delineated boundaries between the soil layers are located at similar depths. Although boring logs at these two CPT testing locations are unavailable, according to the geological setting of the studied site, it seems unlikely that the subsurface stratification configurations and the mechanical properties of subsoils could change significantly within a horizontal clearance smaller than 10 m.

It can be concluded from the above discussion, by fusing multiple CPT records and considering their spatial correlations, the developed joint CPT interpretation algorithm can provide accurate and consistent stratification interpretation results that require barely any additional manual corrections.

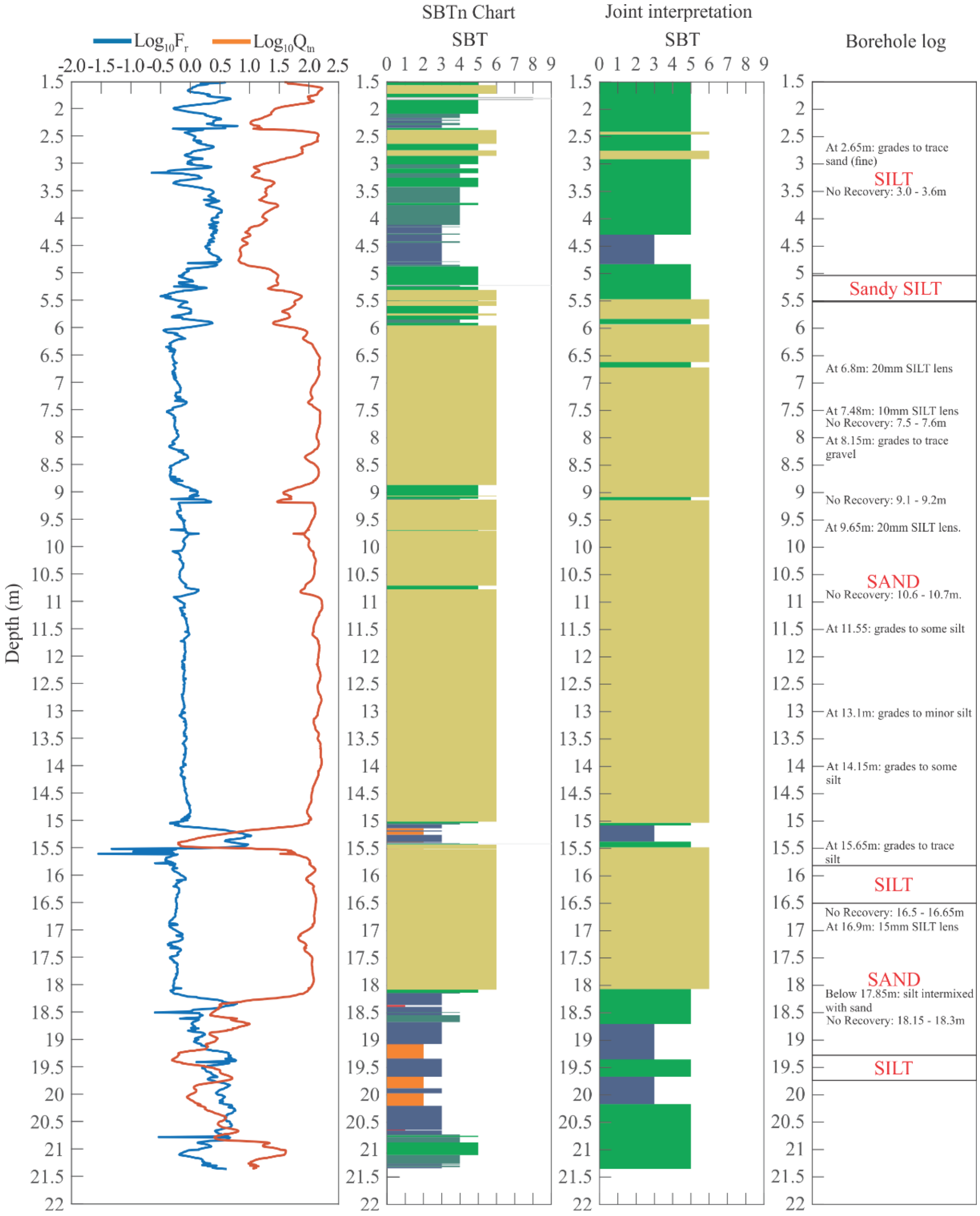


Figure 22. Stratification interpretation results for CPT #6 in the studied Christchurch CBD Area.

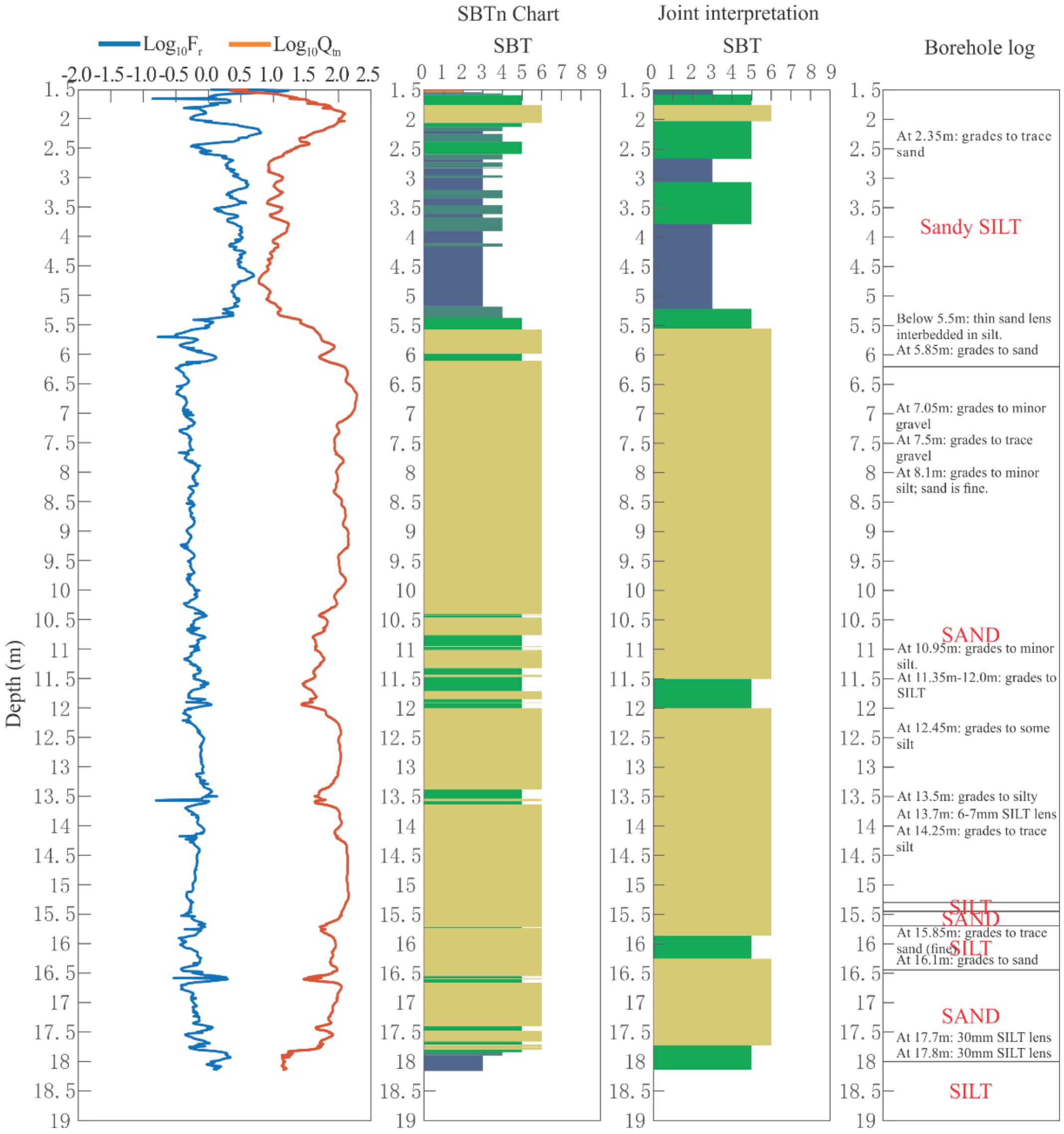


Figure 23. Stratification interpretation results for CPT #7 in the studied Christchurch CBD Area.

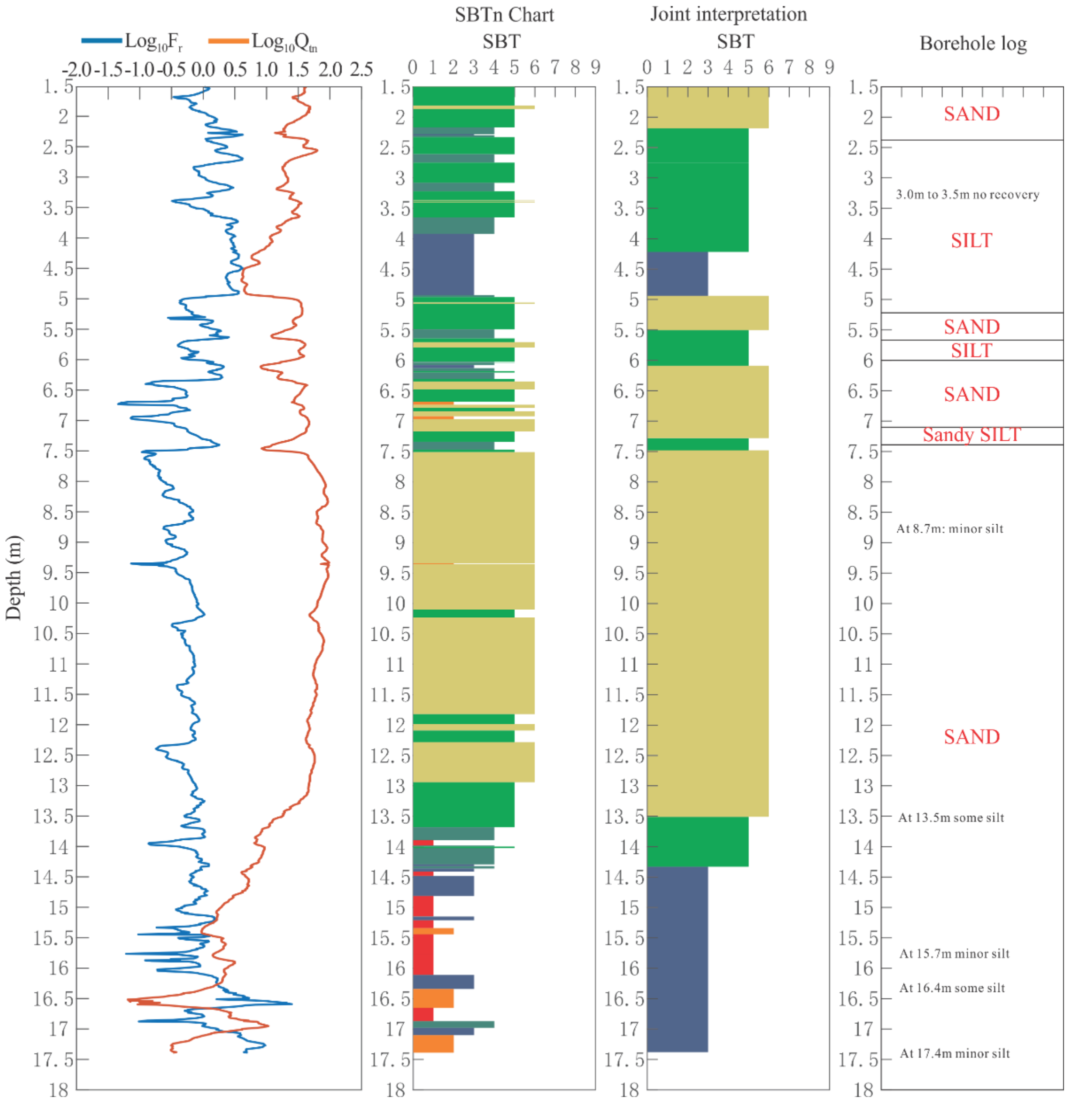


Figure 24. Stratification interpretation results for CPT #12 in the studied Christchurch CBD Area.

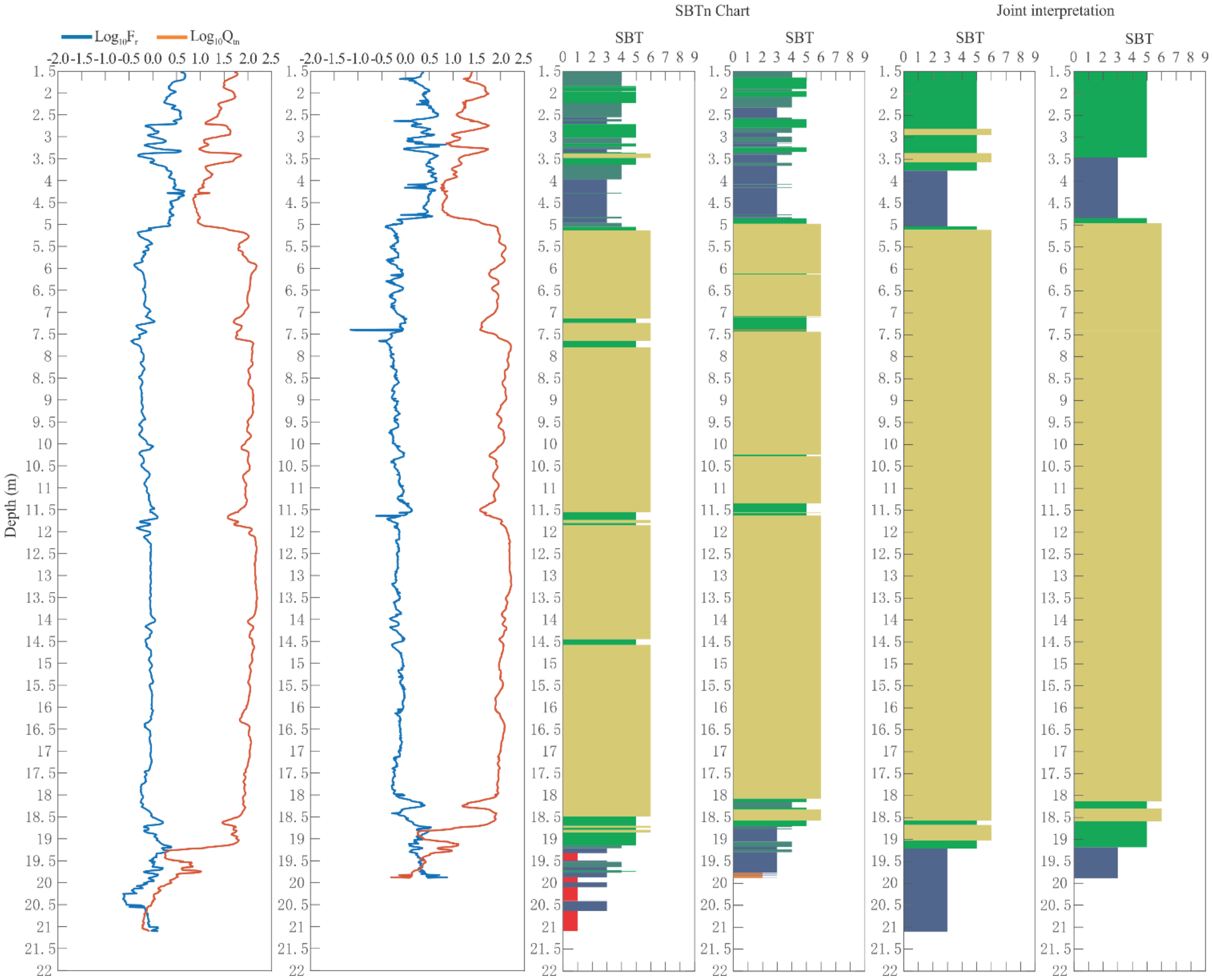


Figure 25. Stratification interpretation results for CPT #1 and # 24 in the studied Christchurch CBD Area.

4. Summary and recommendations

4.1 Major output from this project

This research project report presented a detailed investigation regarding the following three aspects: 1) modifying the developed subsurface modeling program to be able to read DIGGS compliant XML-based schema; 2) Perform multiple case studies using real-world datasets (geotechnical investigation data of several project sites); 3) provide a white paper indicating the potential pathways for making the software accessible for applications by ODOT. The major output from this research project is summarized below:

The DIGGS data schema has been well studied and a python interface has been developed to read and parse the DIGGS .XML files and to construct a customized data structure for the software modules. At this stage, the py-API has been developed in the form of a python module integrated into the developed program package. The py-API can be directly called in a python script and run in a python 3.x console.

Two (2) case studies for validating subsurface profile modeling have been performed. Both DIGGS compliant XML files and archived borehole logs are used for generating a high resolution subsurface profile with quantified uncertainty. The modeling results are discussed and compared with ground truth at selected validation borehole locations. For a more thorough assessment, leave-one-out cross validation also has been conducted. It has been shown that the local modeling accuracy is strongly subjected to the associated uncertainty level at a given pixel. It can be concluded that the uncertainty quantification is paramount in subsurface modeling workflow and the capability of quantifying such uncertainty adds great values to the geotechnical site characterization and downstream engineering practices.

Two (2) case studies for CPT data joint interpretation have been conducted. CPT data from Ohio and New Zealand have been analyzed using the developed joint interpretation module. The stratification interpretation results obtained using the developed joint interpretation algorithm are compared with the corresponding interpretation results obtained using the conventional SBT_n chart. It has been demonstrated that the results obtained using the developed joint interpretation algorithm is much more consistent along both horizontal and vertical directions. The developed joint CPT interpretation algorithm can provide accurate and consistent stratification interpretation results that require barely any additional manual corrections.

4.2 Recommendations for future research

Based on the successful development of the py-API in this project, which enables the data transferring using the geotechnical data format DIGGS between any geotechnical data management systems and the most recent developed AI subsurface modeling tools developed by the research team, now the data-driven modeling paradigm with automatic data flow can be implemented at a research level. However, the high potential of the research outcome from this project in downstream geotechnical practices is far from fully exploited. The research team recommends exploring three possible follow-up research projects in connection with the current project.

(1) Explore the use of geophysical test results to expand current geological modeling capabilities.

Regarding how to incorporate geophysical information into the current subsurface modeling framework, the research team has already explored possible technical route and published two articles (Wang et al. 2019; Wang et al. 2016) in the field of remote sensing and geophysical measurements for characterizing subsurface soil heterogeneity. Further adapting the developed theory and methods to the ODOT needs and aligning it with the geotechnical design purposes can be a feasible path to enhance the current AI based subsurface modeling tool.

(2) Integrate the geological modeling tool into the UASLOPE 3.0 for a total reliability based slope stability analysis and pile stabilization design

For this research direction, the research team has already published several articles in recent years, e.g., Gong et al. (2019) and Wang et al. (2017). The core task for this research topic should be the implementation level rather than theory development level. However, the implementation work may need substantial amount of time and effort to review existing source code of the program (i.e., UASLOPE 3.0), link different computational packages, test additional developments, accommodate exceptions, and develop tutorial and examples.

(3) Enhance the geological modeling tool to possess an ability for direct usage of CPT and/or SPT for estimating pile length.

For this topic, based on the current capability of the joint interpretation using CPT and/or SPT data, embedding the current ODOT foundation design practice of using in-situ tests, such as CPT, SPT, or pressuremeter test results, in conjunction with soil boring logs if applicable, can be integrated into the geological modeling module to allow for greater values and benefits to ODOT engineers and consultants working for ODOT. .

Appendix A: A brief knowledge of the machine learning based subsurface interpretation and modeling framework

A.1 Motivation

A key component of many geological and/or geotechnical subsurface modeling workflows is the use of in-situ testing or borehole log data by expert geologists or geotechnical engineers to build interpretations of the geology (i.e., soil or lithology profile) along cross sections. Since usually only sparse data (i.e., limited data) is available, these interpretations will always be uncertain to some degree. Having a good knowledge of possible variations of subsurface stratifications based on interpretations from incomplete information is important because it allows the end user of a geological model to assess a model's uncertainties and associated confidence level (or accuracy level) as well as to determine how these uncertainties may impact decisions made using the model.

Since geomaterials are natural, rather than manufactured, materials, the planning and design of a built system, a system that is either embedded in or founded on the geomaterials, can be greatly influenced by the site specific soil stratification and variability of soil properties. Therefore, the current practice of interpreting incomplete in-situ testing and/or borehole log data is deficient and inadequate from the viewpoint that individual geologist or engineer exerts his/her own experience and preference in interpretation and that the intrinsic uncertainties of such interpretation cannot be quantified. Due to the nonspecific knowledge of the deposition histories and past tectonic activities, the geological model, a model that characterizes the geological and geotechnical information, at a site may not be known prior to the site investigation. Comprehensive information with 100% coverage could only be achieved by excavating the ground completely, which is obviously not a feasible methodology. Alternatively, geological model is assumed based on limited geotechnical site exploration information and uncertainty of the assumed model is implicitly accepted. However, engineers need to make decision concerning acceptable uncertainty levels of geological models in various stages of planning, design, construction and maintenance of a civil engineering built structure. As a prerequisite for engineers to make a rational decision, there exists a need for developing a methodology to quantitatively assess uncertainties of the interpreted geological model.

It is generally accepted, site investigation plays a vital role in geological and geotechnical practices. Among the various site investigation techniques, cone penetration test and borehole exploration are the most widely used approaches to obtain the subsurface soil information. However, in practice only a limited number of boreholes and/or CPT soundings can be afforded in a specified project, partially due to the limited budget for site investigation and the tight project schedule. As an outcome, the geological and geotechnical information can only be known at relatively sparse borehole and CPT sounding locations; whereas, at other locations these information is unknown and may have to be interpreted based on those from the observed borehole and sounding locations. Under some circumstance, we are even not sure if the number of site exploration locations is sufficient or not for a robust inference of the soil stratification and properties. The incomplete knowledge of the formation process of the geological bodies, together with the possible insufficient number of boreholes, could lead to significant uncertainty in the interpreted geological model. Therefore, it is safe to say that the issues of the characterization of the uncertainty of the interpreted geological model and the influence of such uncertainty on the geotechnical design are long standing challenges to the geologists and engineers.

A.2 Heterogeneity of soil types or geotechnical units

The level of details in establishing spatial distribution of different soil types or geotechnical units depends on specific project needs. Generally, it will be based on a balance between improved details against higher costs. In this project, a simplified soil classification system used by ODOT was used to group all observed soil samples into four types.

Ideally, a detailed and accurate geotechnical analysis relies on obtaining all properties of the soil and rock including all spatial variations of the properties. Obviously, this would be impossible and therefore, a standard procedure is to divide a soil/rock mass into regions of homogeneous geotechnical units. A geotechnical unit is, in theory, a part of the soil or rock mass in which the mechanical properties of the soil or intact rock material are approximately uniform and the mechanical properties of the discontinuities (including anisotropy and spatial correlation of properties) within each set of discontinuities are the same. The anisotropy of properties in a geotechnical unit should be also uniform. In reality, homogeneity is seldom found and material and discontinuity properties vary within a selected range of values within the unit. Therefore, it can be concluded that characterizing the heterogeneity of the geotechnical units is the basic and fundamental task in subsurface modeling.

A.3 Digital subsurface and BIM in geotechnical engineering

The use of Building Information Modeling (BIM) has grown in recent years in structural and infrastructure engineering. However, these BIM software typically ignore information of the geology and subsurface soil properties. This is a significant shortcoming as the whole premise of BIM is to reduce costs by reducing risk at the early design stage and throughout the lifetime of the project. Recently, BIM principles has been applied to geotechnical engineering to help reduce uncertainty and produce a better site investigation practice, which ultimately will help to reduce risk and cost. In the core, BIM concept when extended to geotechnical engineering requires the development of new algorithms in subsurface modeling, transforming from traditional experience-driven paradigm to the more advanced data-driven paradigm, as illustrated in Figure A1. Moreover, instead of interpreting site investigation data in a deterministic manner directly based on engineers' subjective judgement, the raw data can be processed, analyzed and modeled using probabilistic models and Bayesian inferential framework, so that possible spatial patterns can be extracted automatically using AI techniques with the ability to quantify uncertainty. The data-driven paradigm is more objective and the confidence level of the inferred results can be computed using Bayesian machine learning principles. The capability of uncertainty quantification is the key feature that differs the BIM in geotechnical engineering from the counterpart for the above ground structures. In addition, as more and more site investigation data are collected and stored in standardized formats, it is now possible to automatically sharing, transferring, importing, exporting, and processing geotechnical data through internet, which paves the way to developing web-based applications for scalable AI-based subsurface BIM platform.

Digital data is the core and enabler that allows the benefits of BIM to be achieved. For BIM to succeed, common agreed formats for reporting data need to be used. For geotechnical data, the UK and various parts of the world have well established and generally agreed upon standards, as promulgated by AGS (Association of Geotechnical and Geo-environmental Specialists). In the United States, the DIGGS (Data Interchange for Geotechnical and Geoenvironmental Specialists) format, as advocated by ASCE and FHWA, is starting to emerge as the preferred format. Both formats enable the transfer of geotechnical and geo-environmental digital data within and between organizations. When utilizing data interchange standards, compiling geotechnical data requires importing the data from the data interchange file into the chosen geotechnical data management system as illustrated in Figure A2.

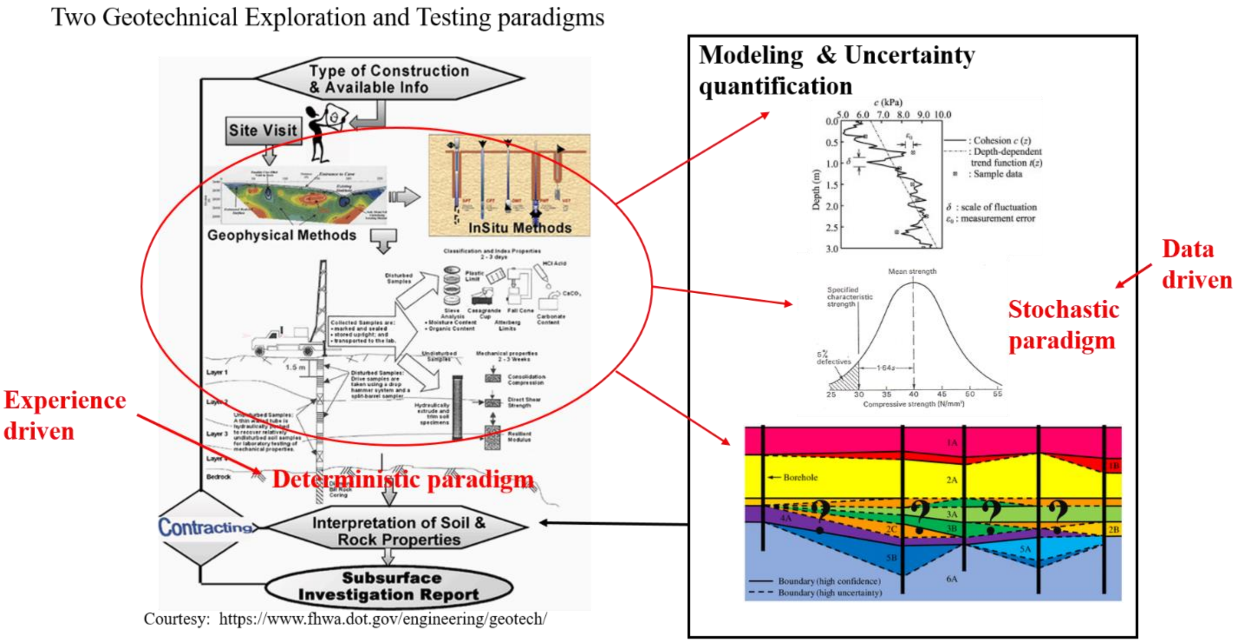

Two Geotechnical Exploration and Testing paradigms
Type of Construction & Available Info
Site Visit
Geophysical Methods
InSitu Methods
Experience driven
Deterministic paradigm
Contracting
Interpretation of Soil & Rock Properties
Subsurface Investigation Report
Courtesy: https://www.fhwa.dot.gov/engineering/geotech/
Modeling & Uncertainty quantification
c (kPa)
Depth (m)
: Cohesion c (z)
: Depth-dependent trend function t(z)
: Sample data
δ : scale of fluctuation
ε₀ : measurement error
Mean strength
Specified characteristic strength
5% defectives
1·64s
Compressive strength (N/mm²)
Stochastic paradigm
Data driven
Borehole
Boundary (high confidence)
Boundary (high uncertainty)

Figure A1. Traditional experience-driven and recent data-driven paradigms of subsurface modeling techniques.

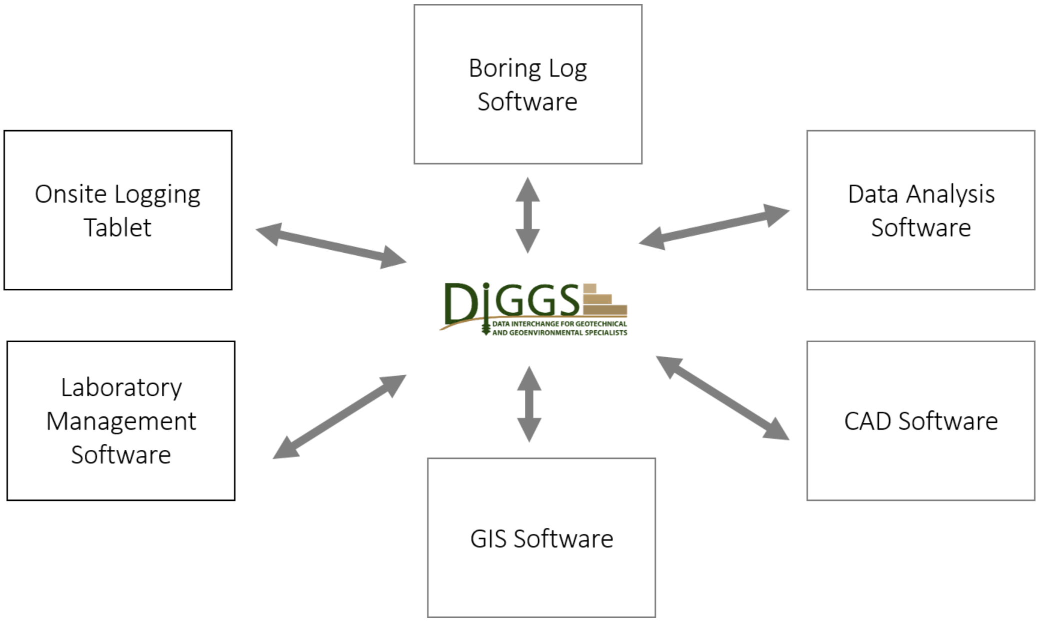

Boring Log Software
Onsite Logging Tablet
Data Analysis Software
DIGGS
DATA INTERCHANGE FOR GEOTECHNICAL AND GEOENVIRONMENTAL SPECIALISTS
Laboratory Management Software
CAD Software
GIS Software

Figure A2. DIGGS schema is the central hub for data sharing and transferring.

A.4 Where is the gap?

1) High resolution (i.e., detailed) 3-D model does not mean uncertainty quantification.

We may have a lot of existing information available for the site, from historic maps, archived plots, to more recent digital data from recent projects. Using all available information would be a sizeable aid in understanding the site geotechnical conditions.

A customary practice could be a two-pronged approach: collating the existing geotechnical knowledge in a geotechnical data management system, and at the same time developing a 3D ground model of the site and ground conditions. The two were then integrated to generate a detailed 3D geotechnical model with the intention of constantly refining the geotechnical data and model throughout the whole process from preliminary investigation, full site investigation, design and onwards. Usually, this kind of 3-D models are considered as the digital twin of the reality, however, with different levels of uncertainties at different locations within the subsurface and at different stages of the project. Even these models are apparently complete and with a lot of details, the portion at unobserved locations is generally considered as deterministic guesses, and there is no associated estimate of confidence level indicating the odds of observing such inference if additional site investigations are conducted. However, such information is critical for decision making and downstream design. Providing an estimate of confidence level of the unobserved parts of the subsurface model is one of major gaps in the current subsurface modeling workflow.

2) Distinguishing two dimensions of uncertainty (epistemic and aleatoric uncertainty) in subsurface modeling.

According to Wikipedia, Uncertainty can be classified into two categories and formally defined as below:

Aleatoric uncertainty: Aleatoric uncertainty is also known as statistical uncertainty, and is representative of unknowns that differ each time when we run the same experiment.

Epistemic uncertainty: Epistemic uncertainty is also known as systematic uncertainty, and is due to things one could in principle know but do not practice. This may be because a measurement is not accurate, because the model neglects certain effects, or because particular data has been deliberately hidden.

In brief, an aleatory concept of uncertainty involves unknown outcomes that can differ each time when one runs an experiment under similar conditions, whereas an epistemic concept of uncertainty involves missing knowledge concerning a fact that either is or is not true. We assert that the philosophical bifurcation of uncertainty mirrors ambivalent intuitions that reside within most decision makers. Returning to the subsurface modeling, the question of whether or not a certain type of soil/rock can be observed within the site could be construed as entailing primarily epistemic uncertainty, whereas the question of the spatial distribution of this type of soil/rock could be construed as entailing primarily aleatory uncertainty. We argue that these forms of uncertainty are not mutually exclusive – indeed, the question of whether the accurate soil/rock profile can be predicted involved both kinds – and that there are systematic planning consequences of whether one form of uncertainty or another is particularly salient to a decision maker. Distinguishing and modeling Epistemic from Aleatory Uncertainty is missing from current subsurface modeling workflow.

3) The way to quantify the uncertainties.

Methodologies for empirically quantifying uncertainties in cross sections derived from borehole data have been developed. In these methodologies, a portion of data are withheld from a set of boreholes that are interpreted by geologists to create cross sections. The withheld data are used to measure the difference between the interpretations and the true geology, as recorded in the withheld borehole. This difference is referred to as the error in the interpretation. The distribution of these errors, and hence the uncertainty, is analyzed statistically to identify factors (such as the local density of boreholes) that determine how the uncertainty behaves. The intention is to see if prediction of uncertainty is possible in future geological settings by using this behavior. Empirical quantification of uncertainty is time consuming and requires dense high-quality data sets, ideally with multiple geologists, to create a range of modeled geologies. In situations where these dense borehole data sets and geologists are not available, there is no proven methodology to follow. From many emerging methods, in recent years, data-driven and AI based methods seem to be promising.

A.5 Bayesian inferential framework and Markov random fields

To cope with the uncertainty involved in the determination of the stratigraphic structures, probabilistic approaches including Bayesian method (Wang et al. 2013), clustering method (Liao and Mayne 2007), wavelet transform modulus maxima method (Ching et al. 2015) and machine learning-based methods

(Wang et al. 2018) have been developed. On the basis of the stratigraphic structures obtained at borehole locations, the stratigraphic configuration at the site could be interpreted from the spatial interpolation of the boundaries between adjacent strata (Chen et al. 2018; Li et al. 2015; Patel and McMechan 2003); the uncertainty in the stratigraphic configuration can be explicitly characterized with the probabilistic approaches that are based upon coupled Markov chain (Elfeki and Dekking 2001; Hu and Huang 2007; Qi et al. 2016) or stochastic Markov random field (Li et al. 2016; Norberg et al. 2002; Wang et al. 2016). To summarize, the uncertainty of the interpreted geological model can be well characterized with the probabilistic approaches, which is composed of two major components: Bayesian machine learning and stochastic model.

1) Bayesian machine learning—the engine of stochastic pattern recognition

Fundamentally, the construction of subsurface model, regardless of being performed by engineers or AI algorithms, is a process of inferring a subsurface model based on local and sparse investigation data. The incompleteness in subsurface sensing/probing data means that the inference should be performed under the guidance of prior knowledge regarding the spatial correlation, structures, and geological histories, and that the uncertainties associated with such inference need to be well quantified. Bayesian machine learning (BML) is an interdisciplinary field between machine learning and Bayesian inferential framework. The former is capable of recognizing abstract patterns and modeling systems with high complexity. The latter enables principled uncertainty estimates, and provides a formal framework for encoding human's prior knowledge into a proper probability model. The selection of this probability model will be of great relevance to the performance and capacity of the developed interpretation and modeling framework. The details will be introduced in the following section. Typically, the subsurface space can be abstracted into a two-level hierarchical model as shown in Figure A3. The first level is referred to as the physical space, which corresponds to the spatial distribution of various types of soils and rocks in the subsurface space. This spatial distribution can be described using a spatial correlated categorical model based on Markov process (i.e., Markov random field). The second level is referred to as the feature space, which represents the variation of the measures (e.g., density, moisture content, permeability, strength, CPT sounding, and SPT N values, etc.) of a single type of subsurface soil, due to the associated inherent heterogeneity. Gaussian assumptions (or log-normal assumptions) can be employed to describe such variation of soil properties. Both of the Markov assumption and Gaussian assumption have long been utilized and validated in existing studies as mentioned above. However, each of them cannot be considered as a comprehensive model for subsurface interpretation and modeling. The Markov assumption alone cannot reflect the variability of soil properties; whereas soil classification only based on Gaussian assumption is vulnerable to the noise of the subsurface data — the interpreted soil stratification is typically unrealistic, manifesting as an excessively frequent change of soil types along depth. Our previous investigations (Wang et al. 2018; Wang et al. 2019; Wang et al. 2018) are the pilot efforts to combine these two assumptions into a hierarchical model, which enables a complete integration of human prior knowledge of the subsurface space with measurable uncertain expression. In addition, from the perspective of model compatibility, such combination is intuitive, as the soil sampling methods and visual/lab inspection provide direct soil label data in the physical space and the in-situ tests and/or geophysical surveys provide measurements of the soil engineering features in the feature space. Therefore, the developed hierarchical model provides a solid underlying framework that enables fusion of diverse subsurface data from different sources and modeled in a unified fashion.

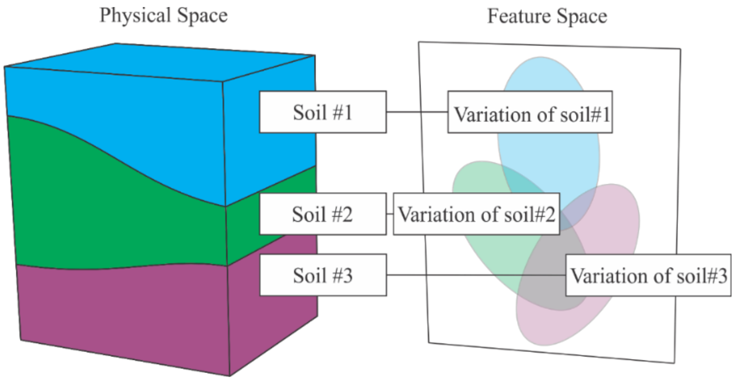


Figure A3 Schematic diagram of the developed hierarchical subsurface conceptual model.

Given the hierarchical subsurface conceptual model as introduced above, the interpretation of subsurface data is essentially a Bayesian inferential classification problem that is based on the similarity of soil property measures in the feature space with consideration of their spatial correlation (aka. spatial pattern) in the physical space.

For an engineering project site with multiple types of subsurface data collected at multiple locations, there are two potential interpretation schemes. The customary scheme is to interpret these data separately, and then manually integrate the obtained local stratification into a complete 3D model. This workflow is subjective and deterministic. The other scheme is to fuse the subsurface data at various locations with multiple data types and interpret them jointly in a unified soil classification system. Our pilot studies demonstrate that the data fusion based joint interpretation scheme has considerable advantages over the conventional interpretation scheme in two aspects. First, from the statistical perspective, additional data samples can provide more complete and enhanced statistical information for each cluster and thereby significantly improve the accuracy of clustering results. Moreover, joint interpretation of subsurface data obtained using multiple site investigation methods can benefit from the “sense of complementary” among different measures; thus, it can further eliminate potential conflicts and lead to consistent interpretation results. More importantly, the capability of joint interpretation serves as the solid assurance of the scalability of the Bayesian inference process, as it provides a reliable way to integrate additional collected subsurface data. To implement such joint interpretation scheme, one typical challenge is to fuse various measures into one single correlation structure while considering various credibility of different measures due to their own measurement mechanisms. This challenge can be addressed by introducing supervised learning or semi-supervised learning algorithms into the unsupervised clustering approaches to honor the correlations among the diverse subsurface measures. The Bayesian inference of the hierarchical model may involve sampling in high-dimensional parameter space for estimating the posterior probability. To enhance efficiency of such inference, Advanced Markov Chain Monte Carlo (MCMC) sampling schemes have been developed to implement the Bayesian inference numerically. Some most recent techniques that mentioned above have already been implemented in the developed subsurface modeling program.

2) Markov random fields—mathematical representations of spatial heterogeneity

The spatial distributions of subsurface soils and the associated engineering features possess certain spatial patterns, which are caused by the natural formation and evolution processes of the subsurface space. Given the localized interpretation results (i.e., the localized subsurface stratification and soil property estimates), the construction of a complete 2D/3D subsurface model requires adequate understanding and effective reproduction of these spatial patterns. To be more specific, the subsurface spatial pattern can be divided into two categories: the stationary pattern and the non-stationary pattern. The former reflects the underlying subsurface features such as the soil composition and the basic texture of their spatial distribution, which are largely determined by the same or similar soil forming processes. The latter is of great relevance to the specific local geological activities, such as land uplift and folding. The real-world subsurface configurations can be considered as an overlaying of the non-stationary localized pattern and the large-scale stationary pattern.

Our preliminary work on remote sensing and geophysical measurements (Wang et al. 2019; Wang et al. 2016) indicate that Markov random fields (MRFs) can be used to mimic the stationary spatial textures and patterns of real-world subsurface. To be more specific, the discretized subsurface can be represented as an undirected graph model in which each soil pixel (in 2-D physical space) or voxel (in 3-D physical space) can be spatially correlated with the nearest neighboring pixels/voxels. This model setting is used to model the basic observed fact that similar soil/rock types are usually close to each other and form layered structures. The anisotropy is usually controlled by setting the parameters in the Markov random field models. Machine learning based pattern recognition algorithms can extract the stationary spatial pattern of subsurface from either continuous geophysics measurements or sparse borehole observations, through a Bayesian inference approach as mentioned above.

For the non-stationary pattern, it can be partially determined by existing geologic knowledge, such as geologic map, if applicable. In most cases that geologic information is missing or insufficient, a straightforward and effective alternative is to perform conditional stochastic simulations based on the obtained stratification results at local points and the developed conceptual model to sample all the possible trends. The non-stationary behavior is controlled by modifying the MRF parameters locally according to existing local prior knowledge.

Once the spatial patterns are extracted, they can then be encoded into proper random field models and reproduced in the generated subsurface realizations. Proper forward stochastic simulation techniques are employed to generate continuous 2D/3D subsurface models. Advanced forward stochastic simulation scheme, such as parallel Gibbs sampler has been implemented in the developed subsurface modeling program to improve the simulation efficiency.

Appendix B: Pathway for converting the existing program into a ready for implementation web-based program for ODOT

At the current stage, the python implementation of the Bayesian machine learning and Markov random field simulator have been well developed. The program packages are ready to use within a python integrated development environment (py-IDE). The DIGGS compliant xml files can be directly read and parsed from hard disk onto high-speed memory and processed by the paralleled programs. After several validation using real-world datasets, the overall performance is satisfying as a prototype in the developing stage.

The next step should be exploiting the pathway for converting the existing python program package into a usable web-based application so that it is robust, easy to use, and free from maintenance at the user end. We will introduce three different strategies.

B.1 Remote SSH based command line interface

Practically every Unix and Linux system includes the ssh command. This command is used to start the SSH client program that enables secure connection to the SSH server on a remote machine. The SSH command is used from logging into the remote machine, transferring files between two machines, and for executing commands on the remote machine. The SSH command provides a secure encrypted connection between two hosts over an insecure network. This connection can also be used for terminal access, file transfers, and for tunneling other applications. Graphical X11 applications for graphical interface can also be run securely over SSH from a remote location.

The developed python program can be deployed on a server with a public web address so that it can be visited and connected by any computer connected into internet. The SSH can be used to encrypt the data traffic between the user end and the server end. The raw data files and the processing executable script file or batch file can be transferred to the server and run at the server. Then the analyzing results can be sent back to the user end via SSH interface.

The advantages of this strategy are listed below:

1) There is no need of additional interface development;

2) Directly using python programming language for flexible usage of the developed program;

3) Independent sever with full software management authority.

The disadvantages include:

1) No graphical interface and hence not user friendly;

2) Raw data needs to be transferred to the server via SSH;

3) A py-IDE is needed at the user-end for editing python scripts.

B.2 Internet based Web Browser User Interfaces

An internet based web browser interface can be developed as a frontend of the developed python program. The key function of this interface will be communicating the user with the backend regarding data transferring and processing. There will be no computation process at the front user end and the only thing needed at the front end is a common web browser. This strategy is highly recommended and the benefits of Web Browser User Interfaces can be summarized as follow:

1) There is no need to manage individual software/program installs and updates whenever there is a change to the software. What's more, if there is a security vulnerability on the network, the only thing needs to be done is update the server and everyone will be updated to a new version all at once. This is a huge time saving benefit of building products that utilize a web browser interface.

2) There is no need to verify that the software will work on various combinations of hardware and operating systems since the web interface is only a communication tool instead of running programs locally. Building a web browser interface means that any customer on any combination of hardware and operating system can access and maximize their use of the developed software.

3) There is no need to worry about compatibility between a graphical user interface and the version of the actual program at the server end.

4) A web browser means that one doesn't have to worry about transferring settings or configuring firewalls—the users can access the software from multiple work locations with ease. When log in, there are no files to transfer—all the work will be saved from the last instance.

5) It is possible to work with multiple tabs open simultaneously for parallel job submitting.

6) The web browser interface can be accessed from phones and tablets. Software is more accessible.

Besides, there are also some long-term value of Web-Based user interface:

First, it's easier to develop web-based user interface compared with other solutions. Developing an independent graphical user interface requires a special set of tools and requires a good knowledge of .net and Java, but it's easier to develop web applications using JavaScript, HTML, and CSS. And when it comes time to test the product, a web-based product allows developers to do more automated testing.

As web services become increasingly prevalent, additional function can become a natural add on when we already have a web server running the product. Building a web-based user interface means that we have laid the foundation for moving toward more web services.

Therefore, building a web-based user interface means that the majority of the users will be accustomed to navigating a web browser. Ultimately, a web-based user interface also allows us to improve the program faster and get new features and versions out to customers at a better pace.

B.3 Python program on a cloud platform

In recent years, software deploying and running are increasingly moving in the cloud, allowing programmers to access and collaborate on their projects on the go and improve the computing performance to some extent. Numerous such services have been launched in the past few years. Serverless computing is one of these services. It is a cloud-computing execution model in which the cloud provider runs the server, and dynamically manages the allocation of machine resources. Pricing is based on the actual amount of resources consumed by an application, rather than on pre-purchased units of capacity. Serverless computing can simplify the process of deploying code into production. Scaling, capacity planning and maintenance operations may be hidden from the developer or operator. Serverless code can be used in conjunction with code deployed in traditional styles. Alternatively, applications can be written to be purely serverless and use no provisioned servers at all.

Google Cloud Platform (GCP) could be a good option to start with. Google Cloud’s serverless platform lets the developers write code their ways without worrying about the underlying infrastructure, deploy functions or apps as source code or as containers, build full stack serverless applications with Google Cloud’s storage, databases, machine learning, and more, and easily extend applications with event-driven computing from Google or third-party service integrations. They can even choose to move their serverless workloads to on-premises environments or to the cloud with great flexibility. Since the source code of the developed python program is ready to use and deploy, GCP infrastructure and computational resources can be leveraged for high performance computing capabilities.

B.4 Steps toward the web-based application

Based on the above information, in this section, a roadmap of the steps toward the web-based application is provided.

Step 1: local server configuration, source code deploying, and testing

As a conservative and simple starting point, a physically standalone computer will be configured and set as a server. The developed source code will be deployed onto this server. A virtual backend will be created and tested. Two types of frontend (SSH and web browser user interface) will be tried. A simple version of the web browser user interface will be developed. The focus will be the speed and quality of data transferring, testing and identifying potential issues, and verify the architecture of the web-based application.

Step 2: Detailed design and development of the web browser user interface

Based on the outcome from Step 1, the frontend will be further modified and refined according to ODOT’s preference and specific requirements so that the user interface can better fit into ODOT current workflow. A beta version will be developed and put it online for testing. Problems and errors will be identified and fixed. This step is straightforward as all the tools and methodologies are well developed and readily available on the market/community.

Step 3: Testing the beta version using several real-world projects

The beta version web browser interface will be tested by the ODOT engineers and staff. Comments and issues will be reported to the developing team. The developing team will maintain and modify the source code of the frontend at the server side. The user can directly access to the updated version via a common web browser.

Step 4: Deploying the program on the cloud computing platform for better scalability and performance

After accumulating enough experiences from Step 1-3 and the overall architecture and workflow are stable. The entire product can be moved to the cloud computing platform for better scalability and performance. Google Cloud Platform can be a good start but several possible options (Amazon AWS, IBM Cloud, etc.) will be compared and evaluated for python compatibility and runtime performance. A cost-benefit analysis will also be carried out as a basis for further commercialization.

References

Chen, G., Zhu, J., Qiang, M., and Gong, W. (2018). "Three-dimensional site characterization with borehole data–a case study of Suzhou area." *Engineering Geology*, 234, 65-82.

Ching, J., Wang, J.-S., Juang, C. H., and Ku, C.-S. (2015). "Cone penetration test (CPT)-based stratigraphic profiling using the wavelet transform modulus maxima method." *Canadian Geotechnical Journal*, 52(12), 1993-2007.

Elfeki, A., and Dekking, M. (2001). "A Markov chain model for subsurface characterization: theory and applications." *Mathematical Geology*, 33(5), 569-589.

Gong, W., Tang, H., Wang, H., Wang, X., and Juang, C. H. (2019). "Probabilistic analysis and design of stabilizing piles in slope considering stratigraphic uncertainty." *Engineering Geology*, 259, 105162.

Hu, Q., and Huang, H. "Risk analysis of soil transition in tunnel works." *Proc., Proceedings of the 33rd ITA-AITES world tunnel congress-underground space-the 4th dimension of metropolises*, 209-215.

Li, X., Zhang, L., and Li, J. (2015). "Using conditioned random field to characterize the variability of geologic profiles." *Journal of Geotechnical and Geoenvironmental Engineering*, 142(4), 04015096.

Li, Z., Wang, X., Wang, H., and Liang, R. Y. (2016). "Quantifying stratigraphic uncertainties by stochastic simulation techniques based on Markov random field." *Engineering Geology*, 201, 106-122.

Liao, T., and Mayne, P. (2007). "Stratigraphic delineation by three-dimensional clustering of piezocone data." *Georisk*, 1(2), 102-119.

Norberg, T., Rosén, L., Baran, A., and Baran, S. (2002). "On modelling discrete geological structures as Markov random fields." *Mathematical Geology*, 34(1), 63-77.

Patel, M. D., and McMechan, G. A. (2003). "Building 2-D stratigraphic and structure models from well log data and control horizons." *Computers & geosciences*, 29(5), 557-567.

Qi, X.-H., Li, D.-Q., Phoon, K.-K., Cao, Z.-J., and Tang, X.-S. (2016). "Simulation of geologic uncertainty using coupled Markov chain." *Engineering geology*, 207, 129-140.

Walthall, S., and Palmer, M. (2006). "The development, implementation and future of the AGS data formats for the transfer of Geotechnical and Geoenvironmental data by electronic means." *GeoCongress 2006: Geotechnical Engineering in the Information Technology Age*, 1-4.

Wang, H., Wang, X., Wellmann, F., and Liang, R. Y. (2018). "A Bayesian unsupervised learning approach for identifying soil stratification using cone penetration data." *Canadian Geotechnical Journal*(ja).

Wang, H., Wellmann, F., Zhang, T., Schaaf, A., Kanig, R. M., Verweij, E., von Hebel, C., and van der Kruk, J. (2019). "Pattern Extraction of Topsoil and Subsoil Heterogeneity and Soil-Crop Interaction Using Unsupervised Bayesian Machine Learning: An Application to Satellite-Derived NDVI Time Series and Electromagnetic Induction Measurements." *Journal of Geophysical Research: Biogeosciences*.

Wang, H., Wellmann, J. F., Li, Z., Wang, X., and Liang, R. Y. (2016). "A Segmentation Approach for Stochastic Geological Modeling Using Hidden Markov Random Fields." *Mathematical Geosciences*, 49(2), 145-177.

Wang, X., Li, Z., Wang, H., Rong, Q., and Liang, R. Y. (2016). "Probabilistic analysis of shield-driven tunnel in multiple strata considering stratigraphic uncertainty." *Structural Safety*, 62, 88-100.

Wang, X., Wang, H., and Liang, R. Y. (2017). "A method for slope stability analysis considering subsurface stratigraphic uncertainty." *Landslides*, 1-12.

Wang, X., Wang, H., Liang, R. Y., and Liu, Y. (2019). "A semi-supervised clustering-based approach for stratification identification using borehole and cone penetration test data." *Engineering Geology*, 248, 102-116.

Wang, X., Wang, H., Liang, R. Y., Zhu, H., and Di, H. (2018). "A hidden Markov random field model based approach for probabilistic site characterization using multiple cone penetration test data." *Structural Safety*, 70, 128-138.

Wang, Y., Huang, K., and Cao, Z. (2013). "Probabilistic identification of underground soil stratification using cone penetration tests." *Canadian Geotechnical Journal*, 50(7), 766-776.

Weaver, S. D., Lefchik, T. E., Hoit, M. I., and Beach, K. (2008). "Geoenvironmental and Geotechnical Data Exchange: Setting the Standard." *GeoCongress 2008: Characterization, Monitoring, and Modeling of GeoSystems*, 557-564.