18º Congresso Brasileiro de Geologia de Engenharia e Ambiental

Data Processing with Computer Vision and Geological Implicit Modelling Application in Viability Study of Eolic Power Plants, a Case Study in Rio Grande do Norte.

R. C. Hammerle [1]; R. Brust Santos [2]; J. Yasbek [3], F. Moura[4]

Resumo – Neste estudo de caso, avaliamos o uso de modelos de Visão Computacional no processamento de dados não tabulados de poços públicos de água subterrânea para criar um modelo geológico em 3D, enfrentando o desafio da baixa densidade de dados de subsuperfície em estudos preliminares de viabilidade de parques eólicos. Python e Tesseract OCR foram usados para extrair dados de poços não estruturados de PDFs do sistema SIAGAS, complementando as 16 perfurações na área de estudo. O fluxo de trabalho permitiu a integração de dados de poços públicos de água subterrânea em um modelo geológico feito em Leapfrog®, em uma área no estado do Rio Grande do Norte, Brasil, melhorando a inferência de subsuperfície para o planejamento de engenharia. Os resultados incluem uma redução de 88% no tempo de processamento, maior densidade de dados geológico-estruturais de 5km² para 1.1km².

Abstract – In this case study, we evaluate the use of Computer Vision models for processing non-tabulated public groundwater well data to create a geological 3D model, addressing the challenge of low subsurface data density in preliminary wind farm feasibility studies. Python and Tesseract OCR were used to extract unstructured well data from SIAGAS system PDFs, supplementing the 16 drilled wells in the study area. The workflow enabled the integration of public groundwater well data into a Leapfrog® -based geological model in an area in Rio Grande do Norte State, Brazil, improving subsurface inference for engineering planning. Results include an 88% reduction in processing time, increased geological-structural data grid density from 5km² to 1.1km²

Palavras-Chave – Transição energética; Visão Computacional; modelagem implícita.

1 Geólogo., ADDA Consultoria: São Paulo – SP, (11) 93472-8279, roberto.hammerle@addaconsultoria.com.br

2 Cientista de Dados Espaciais., MSc, Plandrill: São Paulo - SP, (21) 99778-3383, rodrigobrusts@gmail.com

3 Geólogo., ADDA Consultoria: São Paulo – SP, (11) 99750-0083, julio.yasbek@addaconsultoria.com.br

4 Eng., Msc., Pesquisador: Recife – PE, (81) 9 9197-4617, f.moura.w@gmail.com

1. INTRODUCTION

The global energy transition demands rapid scaling of renewable infrastructure, with wind energy playing a pivotal role due to its near-zero operational greenhouse gas emissions. Brazil, with its vast wind-rich regions like the Rio Grande do Norte State, has emerged as a strategic hub for wind farm development. However, such projects require robust geological feasibility studies to optimize foundation design, mitigate risks (e.g., unstable soils, fractured bedrock), and reduce engineering costs. Traditional approaches rely on sparse geotechnical borehole grids, often resulting in low-resolution 3D geological models that inadequately represent subsurface complexity.

Three-dimensional geological modeling is fundamental for viability assessments, enabling engineers to interpolate subsurface conditions between boreholes. Yet, in preliminary studies, data scarcity persists investigation grids often, leading to grid density cover >=5 km², limiting model accuracy. This study addresses this gap by integrating public groundwater well data from Brazil's Groundwater Information System (SIAGAS). While SIAGAS provides valuable geological profiles, however rendering manual extraction impractical for large-scale projects, especially when it is provided as PDF files.

To overcome these limitations, we developed a computational workflow combining Optical Character Recognition (OCR) and programmatic data processing. By applying Python-based tools and the Tesseract OCR engine, we automated the extraction of lithological and structural data from 183 SIAGAS PDFs, supplementing 16 project-specific boreholes. This approach not only bypassed manual digitization but also enhanced subsurface data density from 5 km² to 1.1 km² per data point. Computer Vision enabled efficient parsing of heterogeneous formats (e.g., driller logs, stratigraphic descriptions), transforming previously unstructured public records into structured inputs for 3D modeling in Leapfrog.

The results demonstrate an 88% reduction in data processing time compared to manual methods, alongside a 4.5x increase in geological-structural data resolution. By integrating AI-enhanced public data, the final model extrapolates subsurface conditions beyond borehole constraints, improving foundation planning accuracy for over 100 wind turbines. This workflow establishes a scalable framework for preliminary feasibility studies in data-sparse regions, aligning cost-effective renewable energy expansion with urgent climate goals.

2. REGIONAL GEOLOGY

Regionally, the study area is located within the geological context of the emerged portion of the Potiguar Basin, specifically in the Potiguar Rift, a structure that is part of the Cretaceous Rift System of Northeast Brazil. The area specifically encompasses parts of the structural features of the Guamaré Graben and Alto Macau and is bounded to the southwest by the Carnaubais Fault, which also limits the Potiguar Rift.

The Guamaré Graben, located in the southeastern portion of the study area, is a linear physiographic depression-oriented NE-SW, with an asymmetric shape and bounded to the southeast by faults that can exceed 5000 meters of displacement. The Macau (horst) Plateau, located in the northwestern portion of the study area, corresponds to the elongated ridges of the basement, arranged parallel to the main axis of the rift (ANP, 2017). Conceptually, the structural context of the study area is illustrated in Figure 1.

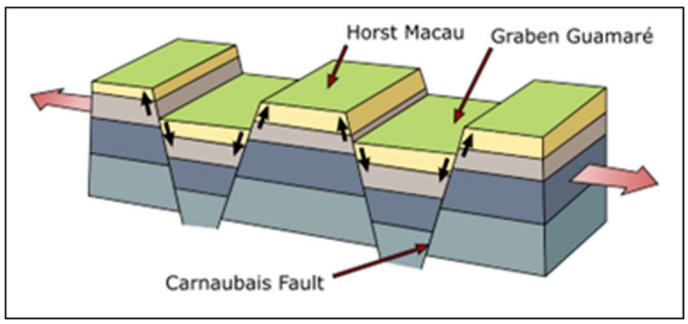

Figure 1. Structural features of the study area.

The basin filling developed according to each phase of its tectonic evolution. Most of these lithostratigraphic units were defined based on the interpretation of the extensive seismic survey and various oil wells present in the Potiguar Basin, which has been the target of oil exploration. In the present study, focusing on surface and outcropping units, the Açu, Jandaíra, Tibau, and Barreiras formations stand out.

The project area for the photovoltaic plant is predominantly situated over the sediments of the Tibau Formation (E3N1t), consisting of fine to conglomeratic sandstones, grey in color, sometimes silicified (Angelim et al., 2006).

3. METHODOLOGY

Fieldwork was carried out between October 2024 and December 2024 and was based on sixteen combined drillings distributed along the lines of the planned aerogenerators.The combined drilling is the investigation method that combines percussion drilling (standard penetration tests) for the soil section and rotary drilling for the rock section. The combined drilling aims to identify, sample, and characterize the entire subsurface substrate, including soils and rocks. The use of motorized equipment allows for the drilling of rock masses and the obtaining of cylindrical-shaped samples, known as drilling cores.

The methodology consists of starting the drilling by performing the SPT until reaching the impenetrable layer (ABNT NBR 6484/2020 and RQF Geotechnical Tests). Upon reaching the impenetrable layer, the continuation of drilling using the rotary method aims to identify and characterize both the upper part of the rock mass, which is generally friable or more altered and fractured, and provide appropriate information about the weathering profile. (NORMA ABGE 104/2023), as well as deeper portions of rock.

Data obtained from field investigations, recorded in physical format on-site, were immediately digitized to ensure information preservation. Lithological descriptions provided by the teams were standardized and unified according to the AGS-BR Guideline SP03/2018. The standardization of NSPT results from different methods and equipment was based on specialized technical literature regarding reference procedures in such cases.

To foster the geotechnical investigation of the initial feasibility study, an extra 182 wells were used in addition to the initial grid of sixteen boreholes. A document containing these digitalized wells, from the database of groundwater extraction wells from the Groundwater Information System (SIAGAS) of the Geological Service of Brazil was used. The processing of the large volume of data was conducted using Tesseract, an OCR (Optical Character Recognition) computer vision algorithm.

Implicit 3D geological-geotechnical modeling for the area designated for the photovoltaic plant was performed using the LeapFrog® software. The methodology developed for the study is summarized in Figure 2.

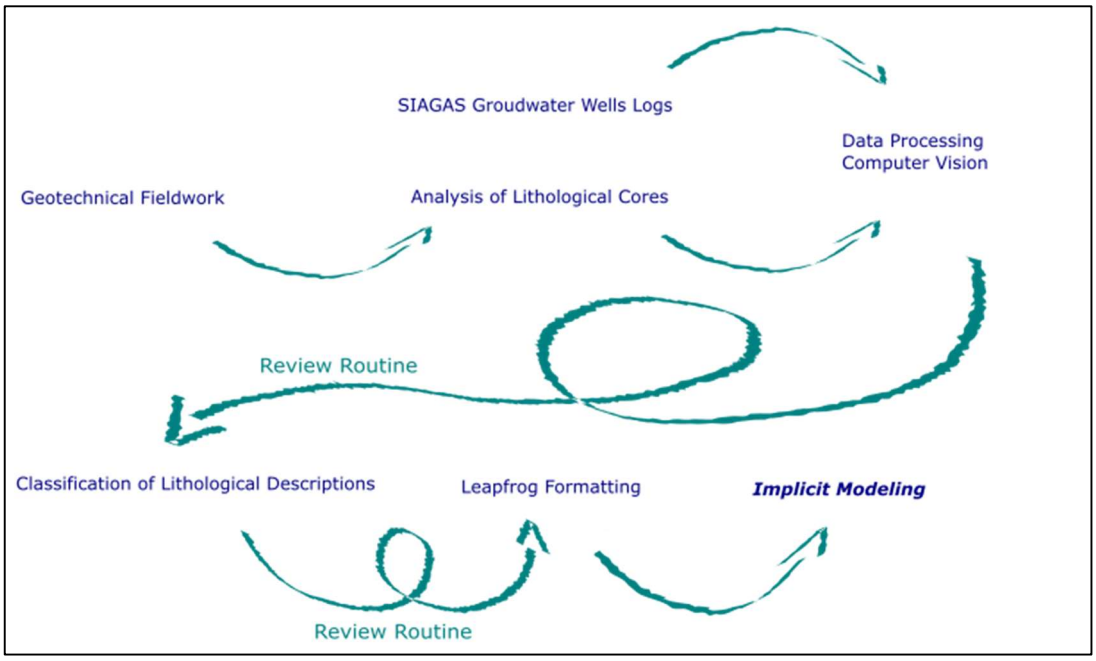

Figure 2. Methodology Developed for the Study.

3.1. Computer Vision Data Processing with OCR Tesseract

Optical Character Recognition (OCR) is a subfield of Computer Vision under the broader discipline of Artificial Intelligence. OCR involves converting textual or handwritten data into machine-readable formats (Awel & Abidi, 2019), enabling automated text and pattern recognition tasks. A widely adopted tool for such applications is Tesseract, an open-source OCR engine initially developed by HP in the 1990s and later maintained by Google (Smith, 2007).

The primary challenge of this study stemmed from the need to extract well log data from 183 PDF documents. Manual extraction would have been prohibitively time-consuming and resource-intensive. To address this, an automated workflow was developed using OCR and Python to efficiently retrieve and format the data into inputs compatible with Leapfrog® geological modeling software.

Each PDF page contained a reference hyperlink to a corresponding groundwater well log page. Exploiting this consistent pattern, a bounding box was programmatically defined around the hyperlink region to extract text via OCR. The extracted text was then stored in a CSV file, enabling users to validate the accuracy and conformity of the logs.

Subsequently, each hyperlink was accessed to retrieve well-specific data embedded in the HTML structure of the linked pages. These pages hosted tabular data detailing well depths, lithological thicknesses, and spatial coordinates. The extracted coordinates facilitated the creation of a geospatial map visualizing well locations, while depth and lithology data informed subsurface geological interpretations.

Finally, a Python-based data processing pipeline was implemented to restructure the tabular data into the Leapfrog® input format. This required generating three distinct datasets: Collar tables (wellhead coordinates), Survey tables (directional surveys), and Interval tables (stratigraphic depth intervals and lithology attributes).

3.2. Implicit 3D Modeling

The implicit 3D modeling methodology used in this study is a particularly useful tool when dealing with complex data sets and aggregated uncertainties, commonly found in mineral exploration and geotechnical engineering. As an alternative to the traditional method of manual interpolations, implicit geological modeling relies on the assistance of mathematical algorithms to infer the geometry and distribution of geological bodies based on the available data.

This methodology allows professionals involved to spend more effort on designing more complex details of the local geology, such as sub-meter lenses of sedimentary units, and the development of various test models from the interpretation of the same data. In addition to these advantages, implicit modeling allows for the automatic update of the model with the addition of new boreholes at any time.

The Leapfrog® software, developed by Seequent, is a consolidated tool in the field of implicit 3D modeling. Leapfrog® uses the FatsRBFTM (Radial Basis Functions) algorithm to fit mathematical models to a large set of 3D data representing complex and non-linear patterns.

The process culminating in implicit modeling is of utmost importance for the quality of the result and consists of the four steps outlined below, which are detailed in the following subsections: 1) contextualizing the local geology using secondary data, such as academic publications and government organization publications; 2) macroscopic (tactile and visual) analysis of the stored core samples and field recognition of the study area; 3) compilation, technical evaluation, and selection of data; and 4) formatting the data to fit the input standards of the modeling.

4. CONCLUSIONS

The data processing with hyperlink extraction using OCR Tesseract took 361 seconds to complete. The most intense processing part was the identification of the hyperlinks and to access the information in HTML files. These two steps were responsible for 97% of the processing time, or around 353 seconds. The other 3% were responsible for formatting, validating and exporting the tables. From all the 183 wells, only 54 had subsurface lithological description, representing 29.5% of the total.

Despite the low rate of well with lithological information, the semi-automated cut significantly the amount of time one would take processing this data manually. If taken into consideration that one well with complete information would take at least 10 minutes to manually process the data, it would mean 540 minutes of underutilization of workforce. The semi-automated workflow could reduce in 88% the processing time, meaning more time for setting the geological model and proper analysis.

By adding the 54 geological profiles from public data sources to the existing 16 boreholes from the project's investigation, the data grid density improved from 5 km² per point to approximately 1.1 km² per point. The benefits in interpreting the local geology are evident in Figure 3.

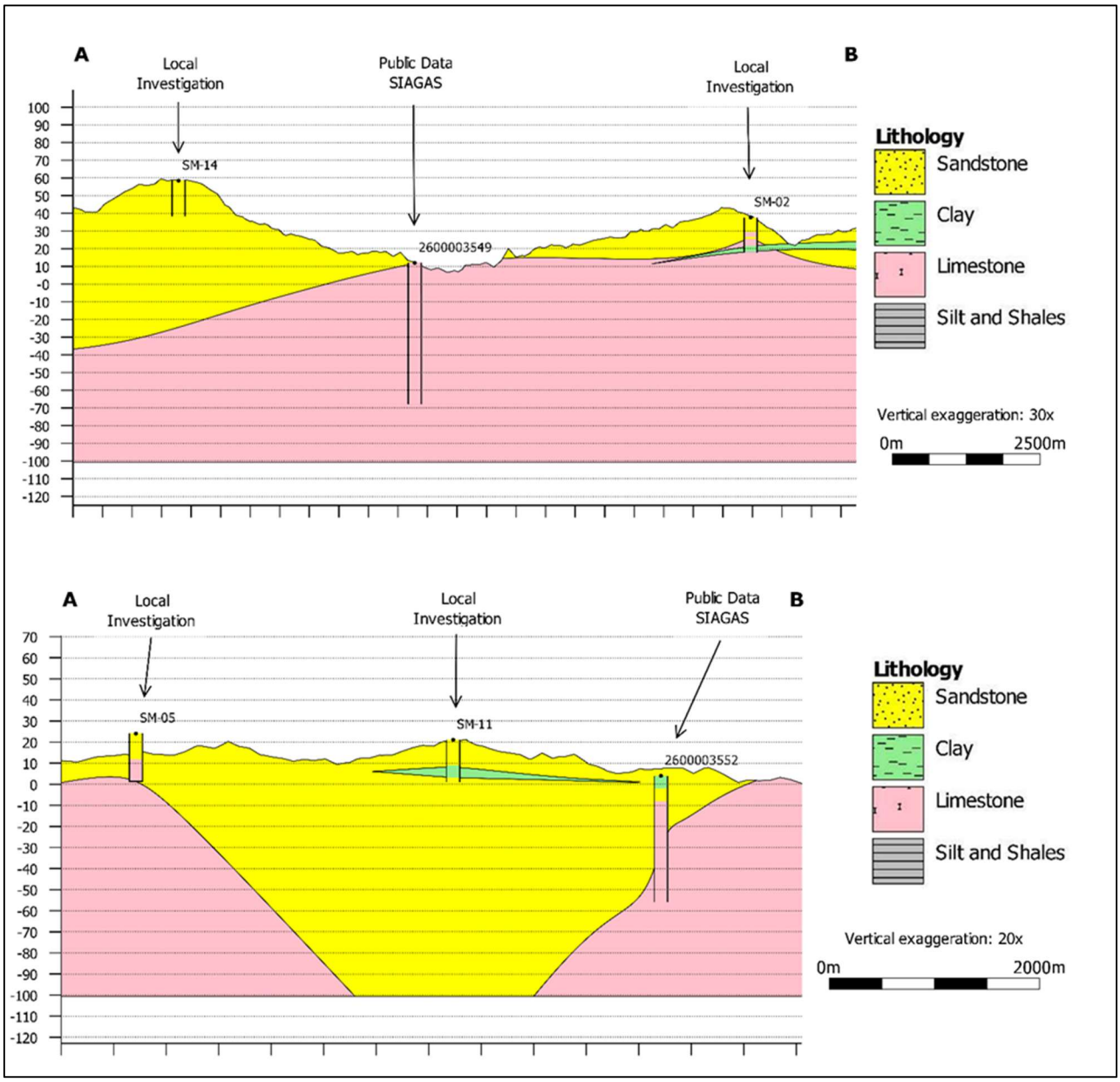

Figure 3 Geological Sections.

REFERÊNCIAS

ABNT NBR 16796. (2020). Solo - Método padrão para avaliação de energia em SPT.

ABNT NBR 6484. (2020). ABNT NBR 6484: Sondagens de Simples Reconhecimento com SPT - Método de Ensaio. Rio de Janeiro.

AWEL, M. A., & ABIDI, A. I. (2019). Review on optical character recognition. International Journal of Research in Engineering and Technology, 6(6), 3666–3669.

ANGELIM et al. (2006). Mapa geológico do Estado do Rio Grande do Norte. Programa Geologia do Brasil – PGB. Projeto Geologia e Recursos Minerais do Estado do Rio Grande do Norte. Recife: CPRM/FAPERN.

ANP. (2017). BACIA POTIGUAR: Sumário Geológico e Setores em Oferta. Superintendência de Definição de Blocos.

Martins, J. B., & Miranda, T. F. (2003). Ensaios de Penetração nos Solos Graníticos da Região Norte de Portugal. Algumas Correlações. Civil Engineering Journal, 17, 5-18.Massad, F. (2005).

NORMA ABGE 104/2023. (s.d.). Sondagem rotativa e sondagem mista. Vários colaboradores. 1a Edição.

SMITH, R., An Overview of the Tesseract OCR Engine. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, 2007, pp. 629-633, DOI: 10.1109/ICDAR.2007.4376991.

SUGUIO, K. (2003). Geologia sedimentar. São Paulo: Edgard Blucher Ltda.