Automated Extraction of Bioclimatic Time Series from PDF Tables

Sabino Maggi, Silvana Fuina and Saverio Vicario, Institute of Atmospheric Research, CNR-IIA, Bari, Italy

Contact: sabino.maggi@cnr.it

European Geosciences Union General Assembly 2013
Vienna 23 – 28 April 2023

Session GI 2.3 | Poster# EGU23-15072

Introduction

Since the development of the original specifications in the 1990s, the PDF document format has become the de facto standard for the distribution and archiving of documents in electronic form due to its ability to preserve the original layout of documents regardless of the hardware, operating system, and application software used to visualize them.

Unfortunately, the PDF format does not contain explicit structural and semantic information, making it very difficult to extract structured information from it, especially data presented in tabular form.

Goal

The automatic extraction of tabular data is a difficult and challenging task, because tables can have extremely different formats and layouts, and involves several complex steps, from the proper recognition and conversion of printed text into machine-encoded characters, to the identification of logically related table constructs (headers, columns, rows, spanning elements), to the breaking down of the data constructs into elemental objects.

Several tools have been developed to support the extraction process. In this work, we survey the most interesting tools for automatic recognition and extraction of tabular data and analyze their advantages and limitations. Special emphasis is given to programmable open source tools because of their flexibility and long-term availability, together with the possibility to easily adapt them to the specific needs of the problem at hand.

As a practical application, we present a workflow based on a set of scripts that can automatically extract daily temperature and precipitation data from the PDF documents made available every year by the Civil Protection of Apulia, Italy.

A selection of available tools

- **pdf2table** (https://www.cvast.tuwien.ac.at/projects/pdf2table) A set of heuristic tools for detecting and decomposing tables in PDF files, storing the extracted data in a structured XML format for easier reuse [1].

- **CELLS** A configurable system for converting unstructured tabular data into a structured form using a set of customizable parameters and ad-hoc heuristics for recovering table cells from text chunks and rulings, generated by the iText4 PDF interpretation library [2, 3].

- **TEXUS** A table-oriented document model framework that uses an abstract table representation that separates the logical structure of the table from its layout. The extraction process produces an canonical abstract representation that can be used by application-specific tasks [4]. The web application built to demonstrate the capabilities of the service is currently not available.

- **GNN-TableExtraction** (https://github.com/AILab-UniFI/GNN-TableExtraction) Redefines the problem as a node classification task and uses a graph neural network to extract table information [5].

- **Sensible** (https://www.sensible.so) An online tool that transforms structured or unstructured documents into data using a description provided by the user. Uses GPT-4 and other large-language models (LLMs) to extract target information.

- **Camelot** (https://camelot-py.readthedocs.io/en/master) Identifies table cells using either whitespace between cells (stream mode) or table borders (lattice mode) and is based on the PDFMiner library to extract individual characters grouped by whitespace into words and sentences.

- **tabula** (https://tabula.technology) A tool similar to Camelot that detects table cells using either a stream mode based on PDFBox or a lattice mode based on OpenCV.

- **pdfplumber** (https://github.com/jsvine/pdfplumber), a library for table extraction and visual debugging.

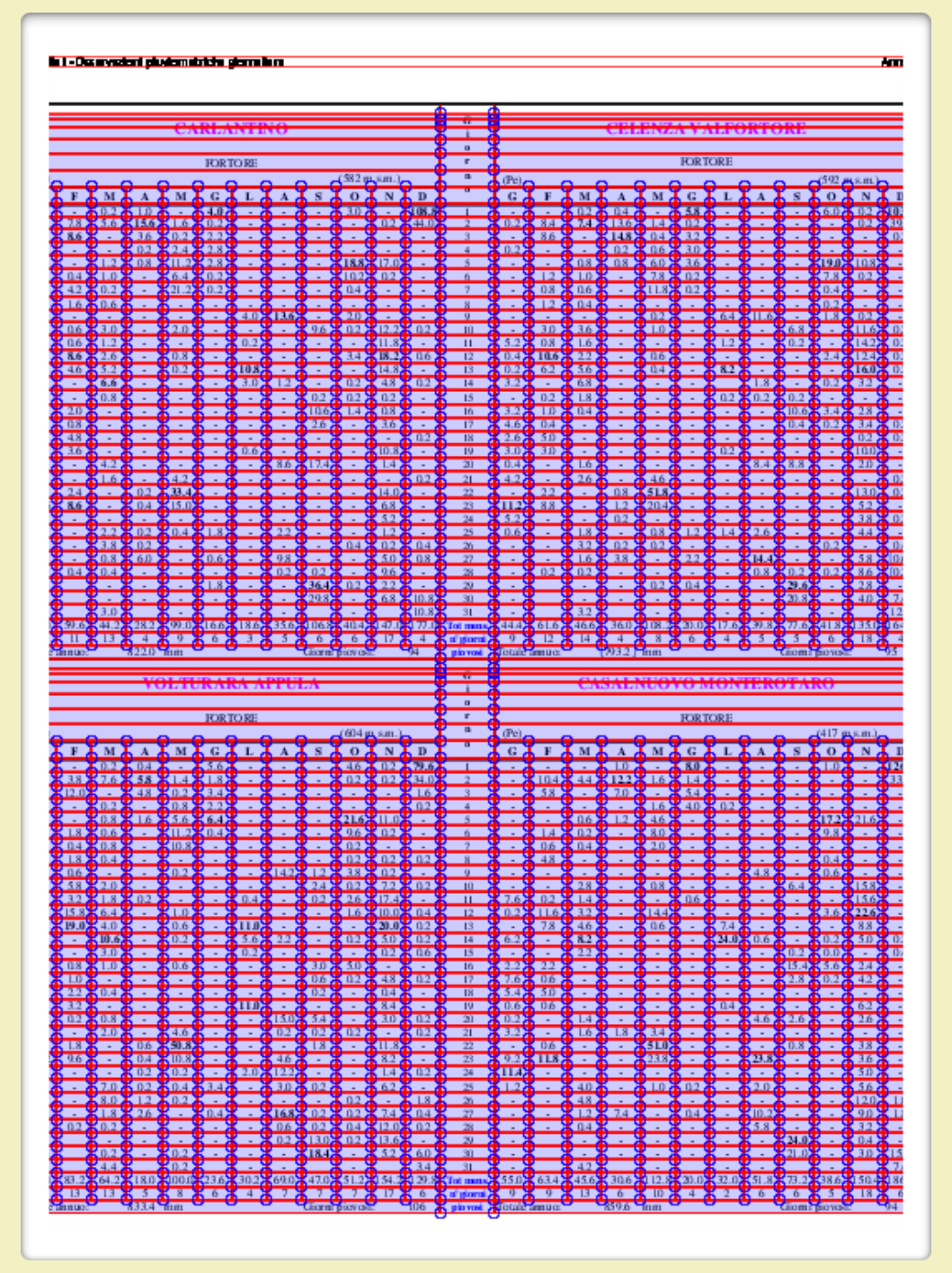
Case study

The data collected by the meteo-hydrometric monitoring network of the Civil Protection of Apulia, Italy, are published annually in PDF format in the so-called Hydrogeologic Annals. The Annals date back to 1922 and represent an invaluable source of information for studying the temporal evolution of the main meteorological variables over the region.

More recent data can be obtained by periodically scraping the data available on the Civil Protection website, a slow and cumbersome process that does not allow obtaining information prior to ~2005.

The PDF files present the measured data as a set of tables with daily measurements for each variable (temperature, precipitation, air pressure, relative humidity, wind, etc.) and for each monitoring station. Each page of the file can contain multiple tables, as shown in the examples below.

To extract temperature and precipitation data from these files we tested several of the available tools, with mixed results. Most of the table contents could be recovered, but all of the tools tested introduced random errors that prevented full extraction of the data. The best results overall were obtained with *pdfplumber*, which however failed to recognize the station names (the purple headers at the top of each table, image on the left), making it impossible to batch process the files.



Workflow and results

A set of custom scripts was developed to perform the recognition and extraction of the temperature and precipitation data stored in the PDF files. The process uses a set of heuristic rules that rely on the knowledge of the layout of the tables to be extracted.

- The full text extraction is performed by an R script based on the pdftools library, which saves the full content of the PDF document as a csv file.
- A set of AWK scripts recognizes the metadata information, the names of the thermometric and pluviometric stations, and the datasets associated with each variable and station.
- A Python script detects errors in the extracted data and allows the user to iteratively correct them.
- A final R script looks up the coordinates associated with each station and converts the extracted data into a time series format, useful for the calculating the derived bioclimatic predictors.

When running the scripts, most of the time is spent on error correction, as the Hydrogeological Annals PDF files are riddled with errors and inconsistencies that make fully automated extraction of the datasets impossible. Some of the most common source of errors are:

- Inconsistent table layout, such as when the header is not preserved across data pages;

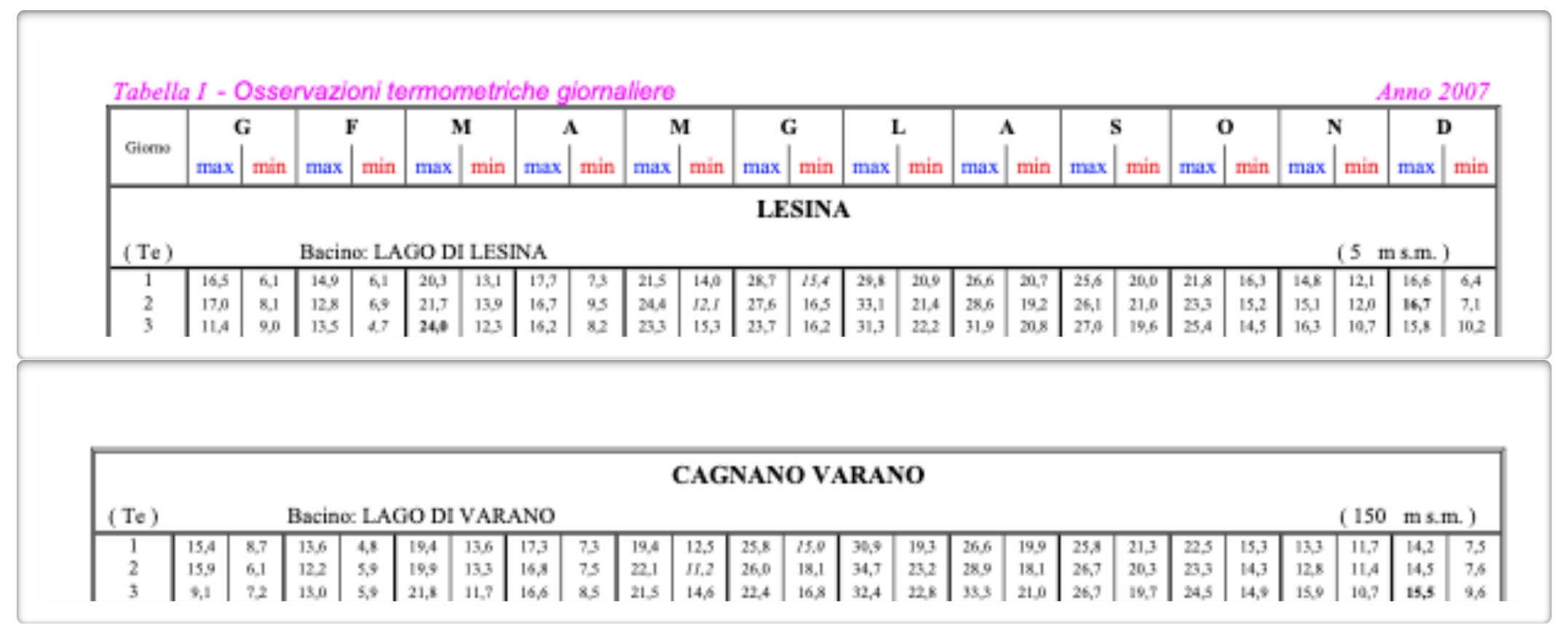


Tabella I - Osservazioni termometriche giornaliere

Anno 2007

LESINA

Bacino: LAGO DI LESINA

CAGNANO VARANO

Bacino: LAGO DI VARANO

- wrong decimal separator (a comma as in plain Italian instead of a period as usual in technical documents);

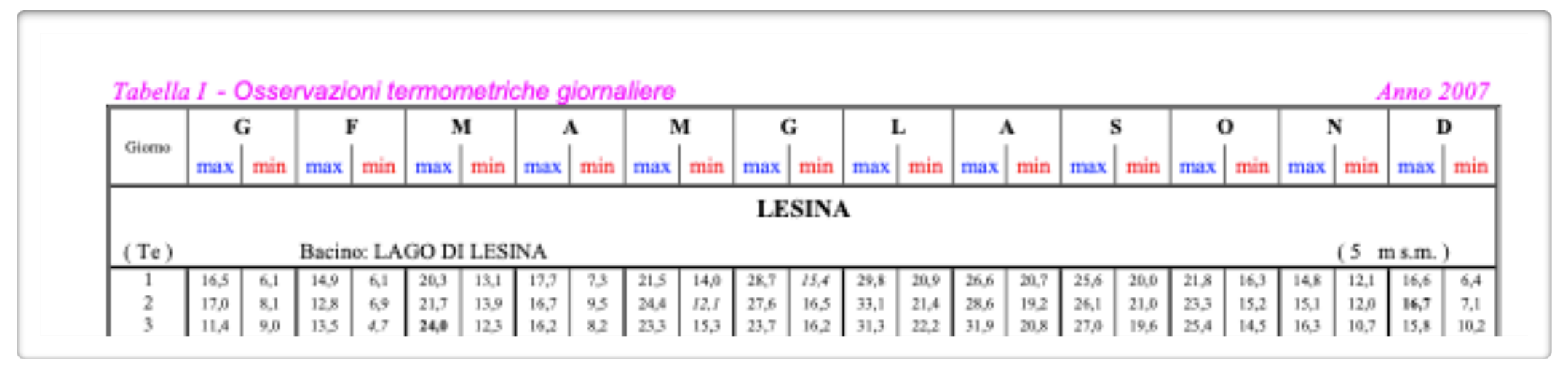


Tabella I - Osservazioni termometriche giornaliere

Anno 2007

LESINA

Bacino: LAGO DI LESINA

- wrong missing/zero data markers;
- missing/zero data markers in non-existent dates;

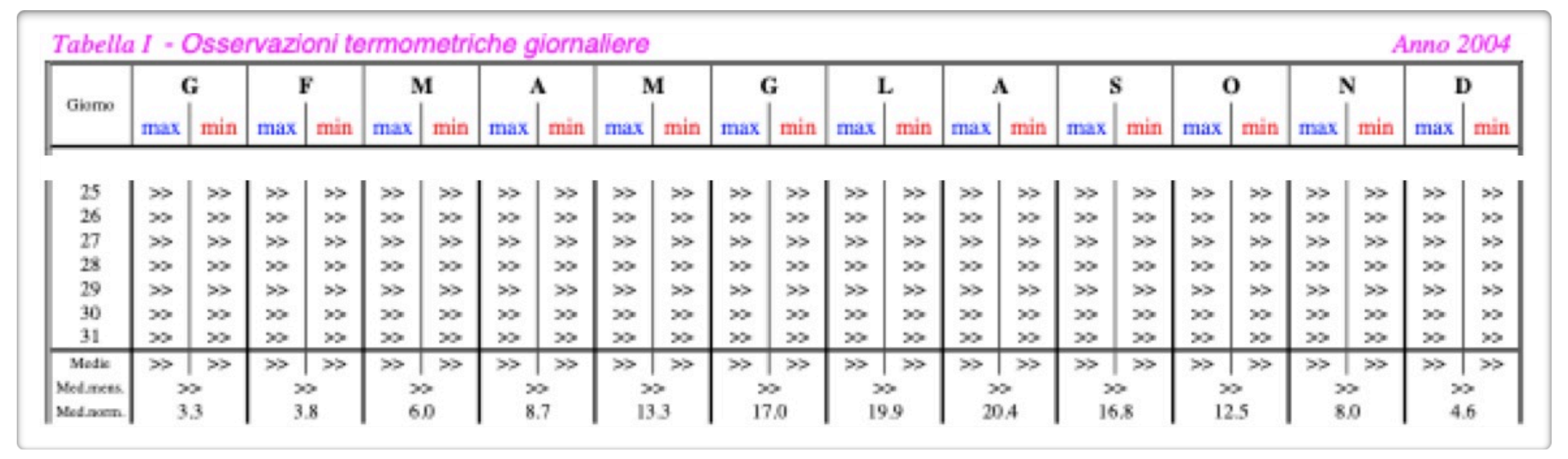
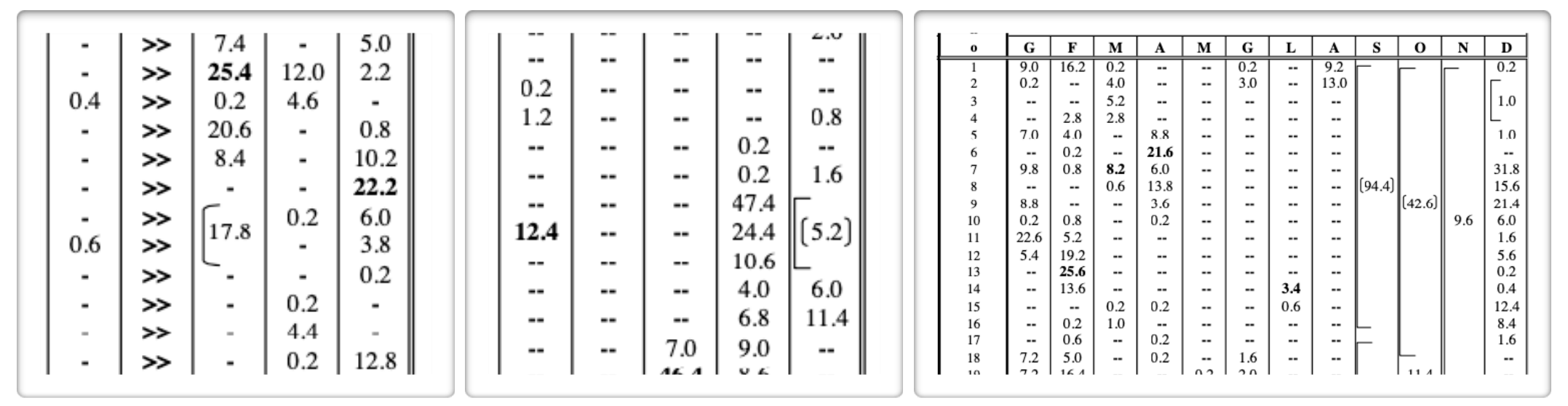


Tabella I - Osservazioni termometriche giornaliere

Anno 2004

- single measurements spanning multiple days;



Currently, the scripts are tailored to extract data embedded in the Hydrogeological Annals of Apulia. The final goal of the project would be to develop a specific table extraction language that could be easily adapted to recognize different types of tables.

References

[1] R. Hastan et al., TEXUS: A unified framework for extracting and understanding tables in PDF documents, Information Processing and Management 56 (2019) 895–918.
[2] A. Shigarov, Table understanding using a rule engine, Expert Systems with Applications 42 (2015) 929–937.
[3] A. Shigarov et al., Configurable Table Structure Recognition in Untagged PDF documents, DocEng '16, September 12-16, 2016, Vienna, Austria.
[4] A. Gemelli et al., Graph Neural Networks and Representation Embedding for Table Extraction in PDF Documents, 2022 26th International Conference on Pattern Recognition (ICPR) August 21-25, 2022, Montréal, Canada.
[5] B Yildiz et al., pdf2table: A Method to Extract Table Information from PDF Files , Proceedings of the 2nd Indian International Conference on Artificial Intelligence (IICAI05), Pune. India. 2005."

https://doi.org/10.5194/egusphere-egu23-15072