A better way to format your document for CEUR-WS

Darren Rawlings[1], Tim Chopard

[1] *University of Groningen, Broerstraat 5, 9712 CP Groningen, Netherlands*

Abstract

A clear and well-documented LaTeX document is presented as an article formatted for publication by CEUR-WS in a conference proceedings. Based on the "ceurart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

Keywords

Multi-Label classification, Principal component analysis, ResNet, Vision Transformers, XGBoost, GeoLifeCLEF 2024, CEUR-WS

1. Introduction

In this research study, we aim to develop a model for predicting plant species in a specific location and time using various environmental factors as predictors. These predictors include satellite images, climatic time series, and other rasterized environmental data such as land cover, human footprint, bioclimatic variables, and soil characteristics. Our motivation behind this challenge is the potential usefulness of accurate plant species prediction in various scenarios related to biodiversity management and conservation, species identification and inventory tools, and education.

We utilized a large-scale training dataset of approximately 5 million plant occurrences in Europe, as well as validation and test sets with over 5,000 and 20,000 plots, respectively. The predicted output will be multi-label, presence-absence data for all present species at each plot. The data covered over 10,000 different plant species, which created significant challendges associated with this task, including learning from single positive labels, dealing with strong class imbalance, multi-modal learning, and handling large-scale datasets.

The potential applications of accurate plant species prediction are numerous. High-resolution maps of species composition and related biodiversity indicators can be created to aid in scientific ecology studies and conservation efforts. The accuracy of species identification tools can be improved by reducing the list of candidate species observable at a given site. Additionally, location-based recommendation services and educational applications with features such as quests or contextualized educational pathways can be developed to facilitate biodiversity inventories and promote environmental education. We

Woodstock'22: Symposium on the irreproducible science, June 07–11, 2022, Woodstock, NY

[†] These authors contributed equally.

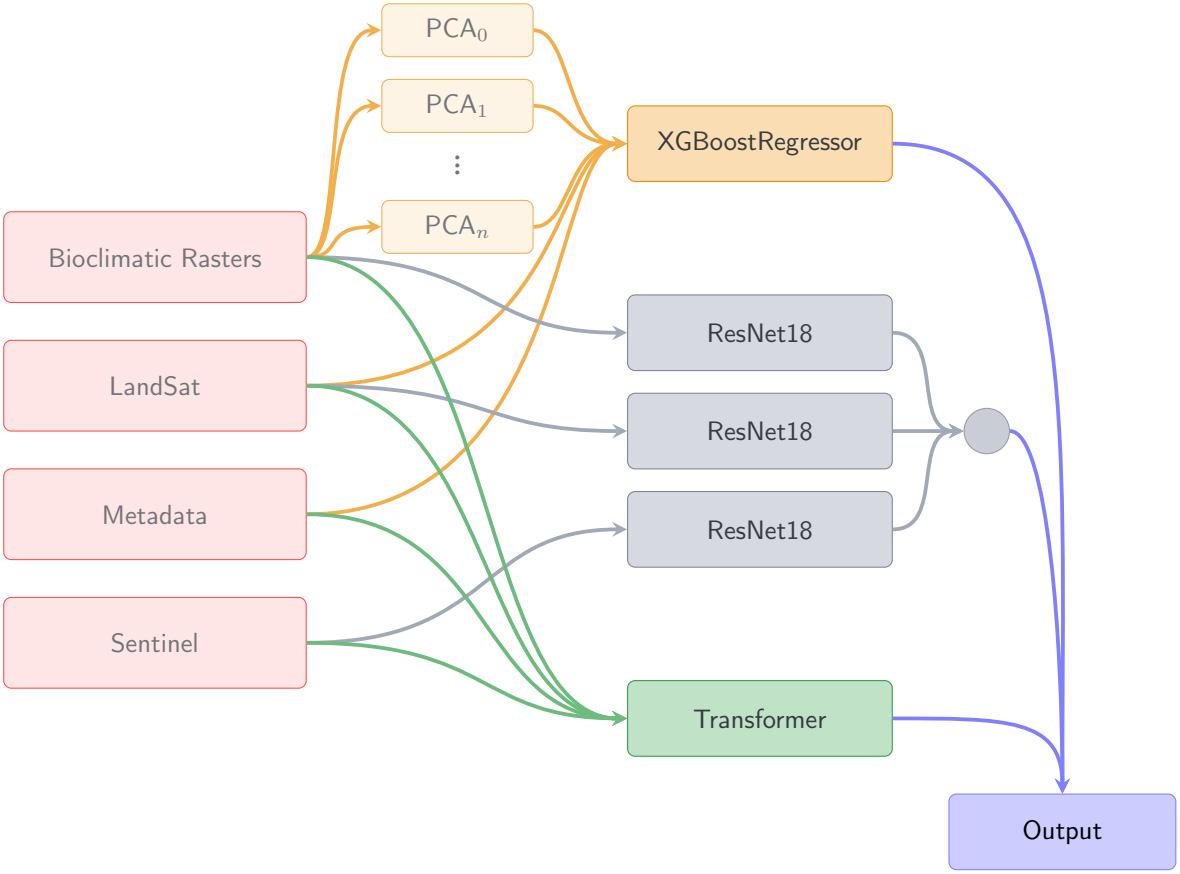

Figure 1: This diagram is just here temporarily and will be in a more relevant section when complete

believe that our research will contribute to the advancement of plant species prediction and its practical applications in various fields.

The research was conducted as part of the GeoLifeCLEF 2024 competition on Kaggle [**?**], which is a part of the LifeCLEF initiative. The competition aims to develop models for predicting plant species in a specific location and time using various environmental factors as predictors.

2. Background

The GeoLifeCLEF challenge has been running for a number of years. Each year, participants are tasked with predicting species distribution, but the challenge has evolved over time, with new datasets, evaluation metrics, and research questions introduced each year. Here, we provide an overview of the some of the recent submissions to GeoLifeCLEF challenge and summarize the key contributions.

In 2021, the GeoLifeCLEF challenge focused on fine-grained visual categorization using remote sensing data. The winning submission by [**?**] leveraged contrastive learning to improve species distribution modeling (SDM) from remote sensing imagery. The authors explored the effectiveness of using only RGB imagery and the impact of adding altitude imagery to the model's performance. They introduced a new consistency-based model selection metric to enhance the model's generalization capabilities. The paper outlined potential areas for further research, including the impact of transformations and the utility of the consistency metric.

In 2022, the GeoLifeCLEF challenge shifted its focus to predicting species distribution across the U.S. and France using remote sensing data and other covariates. The second-place submission by [**?**] proposed a classification approach with a spatial block-label swap regularization during training and an ensemble of deep learning models. Their method achieved a top-30 accuracy of 31.22% on the private test set, securing second place in the competition. The authors reflected on the results and suggested potential improvements and the importance of species distribution modeling for ecological research.

In 2023, the GeoLifeCLEF challenge introduced a new dataset with single positive labels for each location, making multi-label prediction challenging. The winning submission by [**?**] proposed a three-step training strategy to leverage the single positive labels effectively. The authors introduced several CNN-based models and demonstrated their effectiveness compared to a simple baseline. The paper discussed the challenges of the new dataset and the proposed models' performance, providing detailed results and comparisons.

3. Data

Dataset Description

The training data comprises species observations and environmental data. Below, we explain the data in detail.

The species related training data comprises:

Presence-Absence (PA) surveys: including around 90 thousand surveys with roughly 10,000 species of the European flora. The presence-absence data (PA) is provided to compensate for the problem of false-absences of PO data and calibrate models to avoid associated biases. Presence-Only (PO) occurrences: combines around five million observations from numerous datasets gathered from the Global Biodiversity Information Facility (GBIF, www.gbif.org). This data constitutes the larger piece of the training data and covers all countries of our study area, but it has been sampled opportunistically (without standardized sampling protocol), leading to various sampling biases. The local absence of a species among PO data doesn't mean it is truly absent. An observer might not have reported it because it was difficult to "see" it at this time of the year, to identify it as not a monitoring target, or just unattractive.

Environmental data

Besides species data, we provide spatialized geographic and environmental data as additional input variables (see Figure 1). More precisely, For each species observation location, we provide:

Satellite image patches: 3-band (RGB) and 1-band (NIR) 128x128 JPEG images, a color JPEG file for RGB data and a grayscale one for Near-Infrared images at 10m resolution. The source for these images is Sentinel2 remote sensing data pre-processed by the Ecodatacube platform.

Satellite time series: Up to 20 years of values for six satellite bands (R, G, B, NIR, SWIR1, and SWIR2). Each observation is associated with the time series of the satellite median point values over each season since the winter of 1999 for six satellite bands (R, G, B, NIR, SWIR1, and SWIR2). This data carries a high-resolution local signature of the past 20 years' succession of seasonal vegetation changes, potential extreme natural events (fires), or land use changes. The original satellite data has a resolution of 30m per pixel. The source for this is the Landsat remote sensing data pre-processed by the Ecodatacube platform

Environmental rasters Various climatic, pedologic, land use, and human footprint variables at the European scale. We provide scalar values, time-series, and original rasters from which you may extract local 2D images.

Four climatic variables computed monthly (mean, minimum and maximum temperature, and total precipitation) from January 2000 to December 2019, yielding 960 low-resolution (30 arcsec 1 kilometer) rasters covering Europe. The source for these rasters is the CHELSA climate dataset.

Environmental rasters, for each observation, we were provided additional environmental data such as GeoTIFF rasters and scalar values already extracted from the rasters. We provide CSV files, one per band raster type, i.e., Climate, Elevation, Human Footprint, LandCover, and SoilGrids.

Bioclimatic rasters: 19 low-resolution rasters covering Europe; commonly used in species distribution modeling. Provided in longitude/latitude coordinates (WGS84). These were provided as GeoTIFF files with compression and CSV file with extracted values, with a resolution of 30 arcsec (1 kilometer). The source for these rasters is the CHELSA climate dataset.

Soil rasters: Nine pedologic low-resolution rasters covering Europe. Provided variables describe the soil properties from 5 to 15cm depth and are determinant of plant species distributions. Check the definition.txt file about the provided variables (e.g., pH, clay, organic carbon and nitrogen contents, etc.). The format is GeoTIFF files with compression and CSV file with extracted values, with a resolution of 1 kilometer. The source for these rasters is Soilgrids.

Elevation: High-resolution raster covering Europe. Provided as a GeoTIFF file and CSV file with extracted values, with a resolution of 1 arc second (30 meters). The source for this raster is the ASTER Global Digital Elevation Model V3.

Land Cover: A medium-resolution multi-band land cover raster covering Europe. Each band describes either the land cover class prediction or its confidence under various classifications. We recommend the use of IGBP (17 classes) or LCCS (43 classes) layers, often used in species distribution modeling. The format is GeoTIFF file with compression and CSV file with extracted values, with a resolution of 500 meters. The source for this raster is MODIS Terra+Aqua 500m.

Human footprint: Several low-resolution rasters describing human footprint, encap-

sulating seven pressures on the environment (e.g., nighlight level, population density) induced by human presence and activity, are provided for two time periods, the early 90's (1993) and late 2000' (2009). We provide two summary rasters combining all human pressures and two detailed rasters per pressure, which avoid an arbitrary degradation of the original data. The format is GeoTIFF files with compression and CSV file with extracted values, with a resolution of 1 kilometer. The source for these rasters is [**?**].

4. Method

In order to manage the multilabel classification required for this research we implemented an ensemble approach. This section details the individual architectures and the methods used to combine them.

4.1. Architectures

The core architectures used in this project were 18 layer ResNet, XGBoost and [TRANSFORMER]. The outputs of these were then weighted and combined before the maximum arguments were selected.

4.1.1. XGBoost

XGBoost is an open source gradient tree boosting package [**?**]. For this research we used the xgbregression model It has shown broad success on a number of

4.1.2. [Resnet]

4.1.3. [Transformer]

4.2. Process

4.2.1. Species scoring

4.2.2. Species counts

4.3. Metrics

5. Introduction

CEUR-WS's article template provides a consistent LaTeX style for use across CEUR-WS publications, and incorporates accessibility and metadata-extraction functionality. This document will explain the major features of the document class.

If you are new to publishing with CEUR-WS, this document is a valuable guide to the process of preparing your work for publication.

The "`ceurart`" document class can be used to prepare articles for any CEUR-WS publication, and for any stage of publication, from review to final "camera-ready" copy with *very* few changes to the source.

This class depends on the following packages for its proper functioning:

- `natbib.sty` for citation processing;
- `geometry.sty` for margin settings;
- `graphicx.sty` for graphics inclusion;
- `hyperref.sty` optional package if hyperlinking is required in the document;
- `fontawesome5.sty` optional package for bells and whistles.

All the above packages are part of any standard LaTeX installation. Therefore, the users need not be bothered about downloading any extra packages.

6. Modifications

Modifying the template — including but not limited to: adjusting margins, typeface sizes, line spacing, paragraph and list definitions, and the use of the `\vspace` command to manually adjust the vertical spacing between elements of your work — is not allowed.

7. Template parameters

There are a number of template parameters which modify some part of the `ceurart` document class. This parameters are enclosed in square brackets and are a part of the `\documentclass` command:

```
\documentclass[parameter]{ceurart}
```

Frequently-used parameters, or combinations of parameters, include:

- `twocolumn` : Two column layout.
- `hf` : Enable header and footer[1].

8. Front matter

8.1. Title Information

The titles of papers should be either all use the emphasizing capitalized style or they should all use the regular English (or native language) style. It does not make a good impression if you or your authors mix the styles.

Use the `\title` command to define the title of your work. Do not insert line breaks in your title.

8.2. Title variants

`\title` command have the below options:

- `title`: Document title. This is default option.

  ```
  \title[mode=title]{This is a title}
  ```

  You can just omit it, like as follows:

  ```
  \title{This is a title}
  ```

- `alt`: Alternate title.

[1]You can enable the display of page numbers in the final version of the entire collection. In this case, you should adhere to the end-to-end pagination of individual papers.

```
\title[mode=alt]{This is a alternate title}
```

- `sub`: Sub title.

  ```
  \title[mode=sub]{This is a sub title}
  ```

  You can just use `\subtitle` command, as follows:

  ```
  \subtitle{This is a sub title}
  ```

- `trans`: Translated title.

  ```
  \title[mode=trans]{This is a translated title}
  ```

- `transsub`: Translated sub title.

  ```
  \title[mode=transsub]{This is a translated sub title}
  ```

8.3. Authors and Affiliations

Each author must be defined separately for accurate metadata identification. Multiple authors may share one affiliation. Authors' names should not be abbreviated; use full first names wherever possible. Include authors' e-mail addresses whenever possible.

`\author` command have the below options:

- `style` : Style of author name (chinese)
- `prefix` : Prefix
- `suffix` : Suffix
- `degree` : Degree
- `role` : Role
- `orcid` : ORCID
- `email` : E-mail
- `url` : URL

Author names can have some kinds of marks and notes:

- affiliation mark: `\author[<num>]`.

The author names and affiliations could be formatted in two ways:

1. Group the authors per affiliation.
2. Use an explicit mark to indicate the affiliations.

Author block example:

```
\author[1,2]{Author Name}[%
    prefix=Prof.,
    degree=D.Sc.,
    role=Researcher,
    orcid=0000-0000-000-0000,
    email=name@example.com,
    url=https://name.example.com
]

\address[1]{Affiliation #1}
\address[2]{Affiliation #2}
```

8.4. Abstract and Keywords

Abstract shall be entered in an environment that starts with `\begin{abstract}` and ends with `\end{abstract}`.

```
\begin{abstract}
  This is an abstract.
\end{abstract}
```

The key words are enclosed in a `keywords` environment. Use `\sep` to separate keywords.

```
\begin{keywords}
  First keyword \sep
  Second keyword \sep
  Third keyword \sep
  Fourth keyword
\end{keywords}
```

At the end of front matter add `\maketitle` command.

8.5. Various Marks in the Front Matter

The front matter becomes complicated due to various kinds of notes and marks to the title and author names. Marks in the title will be denoted by a star ($\star$) mark; footnotes are denoted by super scripted Arabic numerals, corresponding author by an Conformal asterisk (*) mark.

8.5.1. Title marks

Title mark can be entered by the command, `\tnotemark[<num>]` and the corresponding text can be entered with the command `\tnotetext[<num>]{<text>}`. An example will be:

```
\title{A better way to format your document for CEUR-WS}

\tnotemark[1]
\tnotetext[1]{You can use this document as the template for preparing
↪  your
  publication. We recommend using the latest version of the ceurart
  ↪  style.}
```

`\tnotemark` and `\tnotetext` can be anywhere in the front matter, but should be before `\maketitle` command.

8.5.2. Author marks

Author names can have some kinds of marks and notes:

- footnote mark : `\fnmark[<num>]`
- footnote text : `\fntext[<num>]{<text>}`
- corresponding author mark : `\cormark[<num>]`
- corresponding author text : `\cortext[<num>]{<text>}`

8.5.3. Other marks

At times, authors want footnotes which leave no marks in the author names. The note text shall be listed as part of the front matter notes. Class files provides `\nonumnote` for this purpose. The usage

```
\nonumnote{<text>}
```

and should be entered anywhere before the `\maketitle` command for this to take effect.

9. Sectioning Commands

Your work should use standard LaTeX sectioning commands: `\section`, `\subsection`, `\subsubsection`, and `\paragraph`. They should be numbered; do not remove the numbering from the commands.

Simulating a sectioning command by setting the first word or words of a paragraph in boldface or italicized text is not allowed.

10. Tables

The "`ceurart`" document class includes the "`booktabs`" package — https://ctan.org/pkg/booktabs — for preparing high-quality tables.

Table captions are placed *above* the table.

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper "floating" placement of tables, use the environment `table` to enclose the table's contents and the table caption. The contents of the table itself must go in the `tabular` environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules.

Immediately following this sentence is the point at which Table **??** is included in the input file; compare the placement of the table here with the table in the printed output of this document.

To set a wider table, which takes up the whole width of the page's live area, use the environment `table*` to enclose the table's contents and the table caption. As with a single-column table, this wide table will "float" to a location deemed more desirable. Immediately following this sentence is the point at which Table **??** is included in the input file; again, it is instructive to compare the placement of the table here with the table in the printed output of this document.

Table 1
Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
\$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

Table 2
Some Typical Commands

Command	A Number	Comments
`\author`	100	Author
`\table`	300	For tables
`\table*`	400	For wider tables

11. Math Equations

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. Each of the three are discussed in the next sections.

11.1. Inline (In-text) Equations

A formula that appears in the running text is called an inline or in-text formula. It is produced by the `math` environment, which can be invoked with the usual `\begin` ... `\end` construction or with the short form `$` ... `$`. You can use any of the symbols and structures, from α to ω, available in LaTeX [**?**]; this section will simply show a few examples of in-text equations in context. Notice how this equation: $\lim_{n\to\infty} \frac{1}{n} = 0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

11.2. Display Equations

A numbered display equation—one set off by vertical space from the text and centered horizontally—is produced by the `equation` environment. An unnumbered display equation is produced by the `displaymath` environment.

Again, in either environment, you can use any of the symbols and structures available in LaTeX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n\to\infty} \frac{1}{n} = 0. \tag{1}$$

Notice how it is formatted somewhat differently in the `displaymath` environment. Now,

we'll enter an unnumbered equation:

$$S_n = \sum_{i=1}^{n} x_i,$$

and follow it with another numbered equation:

$$\lim_{x \to 0} (1 + x)^{1/x} = e \tag{2}$$

just to demonstrate LaTeX's able handling of numbering.

12. Figures

The "`figure`" environment should be used for figures. One or more images can be placed within a figure. If your figure contains third-party material, you must clearly identify it as such, as shown in the example below.

Your figures should contain a caption which describes the figure to the reader. Figure captions go below the figure. Your figures should also include a description suitable for screen readers, to assist the visually-challenged to better understand your work.

Figure captions are placed below the figure.

13. Citations and Bibliographies

The use of BibTeX for the preparation and formatting of one's references is strongly recommended. Authors' names should be complete — use full first names ("Donald E. Knuth") not initials ("D. E. Knuth") — and the salient identifying features of a reference should be included: title, year, volume, number, pages, article DOI, etc.

The bibliography is included in your source document with these two commands, placed just before the `\end{document}` command:

```
\bibliography{bibfile}
```

where "`bibfile`" is the name, without the "`.bib`" suffix, of the BibTeX file.

13.1. Some examples

A paginated journal article [**?**], an enumerated journal article [**?**], a reference to an entire issue [**?**], a monograph (whole book) [**?**], a monograph/whole book in a series (see 2a in spec. document) [**?**], a divisible-book such as an anthology or compilation [**?**] followed by the same example, however we only output the series if the volume number is given [**?**] (so series should not be present since it has no vol. no.), a chapter in a divisible book [**?**], a chapter in a divisible book in a series [**?**], a multi-volume work as book [**?**], an article in a proceedings (of a conference, symposium, workshop for example) (paginated proceedings article) [**?**], a proceedings article with all possible elements [**?**], an example of an enumerated proceedings article [**?**], an informally published work [**?**],

Figure 2: 1907 Franklin Model D roadster. Photograph by Harris & Ewing, Inc. [Public domain], via Wikimedia Commons. (https://goo.gl/VLCRBB).

a doctoral dissertation [**?**], a master's thesis: [**?**], an online document / world wide web resource [**?** **?** **?**], a video game (Case 1) [**?**] and (Case 2) [**?**] and [**?**] and (Case 3) a patent [**?**], work accepted for publication [**?**], prolific author [**?**] and [**?**]. Other cites might contain 'duplicate' DOI and URLs (some SIAM articles) [**?**]. Multi-volume works as books [**?**] and [**?**]. A couple of citations with DOIs: [**?** **?**]. Online citations: [**?** **?** **?** **?**].

14. Acknowledgments

Identification of funding sources and other support, and thanks to individuals and groups that assisted in the research and the preparation of the work should be included in an acknowledgment section, which is placed just before the reference section in your document.

This section has a special environment:

```
\begin{acknowledgments}
  These are different acknowledgments.
\end{acknowledgments}
```

so that the information contained therein can be more easily collected during the article metadata extraction phase, and to ensure consistency in the spelling of the section heading.

Authors should not prepare this section as a numbered or unnumbered `\section`; please use the "`acknowledgments`" environment.

15. Appendices

If your work needs an appendix, add it before the "`\end{document}`" command at the conclusion of your source document.

Start the appendix with the "`\appendix`" command:

```
\appendix
```

and note that in the appendix, sections are lettered, not numbered.

Acknowledgments

Thanks to the developers of ACM consolidated LaTeX styles https://github.com/borisveytsman/acmart and to the developers of Elsevier updated LaTeX templates https://www.ctan.org/tex-archive/macros/latex/contrib/els-cas-templates.

A. Online Resources

The sources for the ceur-art style are available via

- GitHub,
- Overleaf template.