ADVANCES IN FOREST FIRE RESEARCH

2022

Edited by
DOMINGOS XAVIER VIEGAS
LUÍS MÁRIO RIBEIRO

Creating a forest disturbance dataset for continental Portugal

Eduardo Fernandes*; Carlos Viegas Damásio; João Moura Pires

NOVA LINCS and Departamento de Informática of NOVA School of Science and Technology.
Quinta da Torre, 2829-516 Caparica, Portugal,
{efa.fernandes@campus.fct.unl.pt}, {cd, jmp}@fct.unl.pt

**Corresponding author*

Keywords

Dataset, forest disturbance, changepoint detection, time series, Landsat

Abstract

Quality data is of the utmost importance to effectively evaluate any classification technique. In the realm of remote sensing, more precisely forest disturbance, different studies use different custom datasets. This paper shows the methodology used to create a forest disturbance dataset for continental Portugal. The dataset contains 664 forest points generated from a stratified random sampling based on the tree species and climate zone covering continental Portugal. Every point contains the auxiliary data used and a thorough list of all years where a disturbance occurs and the respective reason for the disturbance; the years span from 1986 until 2019.

The analysis was performed with the Google Earth Engine platform for a fast and flexible solution, tailored to the dataset's needs. To complement the satellite time series, five auxiliary datasets were used to understand the cause of the disturbances and increase confidence.

The resulting dataset shows known differences among various parts of the county. While the north area of the study contains a bigger number of points with multiple disturbances, the south is exactly the opposite with an abundance of undisturbed points. These asymmetries are also reflected in the species present in these different regions.

This dataset may be of interest for studies that need to evaluate forest disturbance techniques, or even changepoint detection techniques, and it may also be useful in studies that focus on differentiating types of forest recovery.

The dataset is available on GitHub at https://github.com/EduardoFAFernandes/portuguese-forest-disturbance-dataset/ 3

1. Introduction

In the remote sensing community, there have been studies that propose or improve disturbance detection techniques, either covering all types of land-covers (Kennedy et al., 2010; Zhu & Woodcock, 2014) or focusing only on forests areas (Brooks et al., 2014; Vogelmann et al., 2012). All face a common challenge: evaluating their approaches. Some make a qualitative evaluation (Vogelmann et al., 2012) while others strive for a quantitative evaluation that inherently needs solid reference data (Brooks et al., 2014; Huang et al., 2010; Vogelmann et al., 2012; Zhu & Woodcock, 2014). Obtaining this type of data is not easy and the aforementioned studies even felt the need to create their datasets. TimeSync (Cohen et al., 2010) was created for this purpose, it is a tool to visualise Landsat time series and collect annotated data; this tool is used in (Kennedy et al., 2010) and (Cohen et al., 2017) to generate disturbance datasets for the USA. We have created our tool to do this type of analysis, the reason will be explained in **Section 2.3**.

This paper will describe the creation of a forest disturbance dataset for continental Portugal and perform a small analysis and interpretation of the collected data.

This work was partially supported by Fundação para a Ciência e a Tecnologia (FCT.IP) through project Floresta Limpa (PCIF/MOG/0161/2019) and NOVA LINCS (UIDB/04516/2020).

2. Materials and Methods

To create the dataset three main steps were taken. The first step was to select points to be analysed. The second step was to gather auxiliary information like fires and land cover data to support evaluation, and to create a sample that would be representative of continental Portugal. The third and final step consisted of individually validating and analysing every point using Landsat time series, auxiliary information, and historical images from Google Earth Engine (GEE) (Gorelick et al., 2017).

2.1. Points to Evaluate

To select only points from forest areas two data sources were used: Inventário Florestal Nacional (IFN) the Portuguese forest inventory and a custom climate region dataset created with feedback from IPMA. To know what species were present at every point we sampled the latest available IFN at the time (2015), using a stratified random sampling based on the species and the climate region. For the species stratification, the classes in IFN 2015 were used, namely: Acacias, Carob Tree, Chestnut, Cork Oak, Eucalyptus, Holm Oak, Maritime Pine, Oak, Other Hardwoods, Other Softwoods, and Stone Pine. Regarding the climate region, five distinct regions were used: North Coastal, North Interior, Centre Coastal, Centre Interior, Alentejo and Algarve.

2.2. Auxiliary Information

Four official data sources were considered relevant for the creation of the dataset describing land cover, species, and burned areas (see **Table 1** and **Figure 1**), the datasets, have been regularly updated throughout the years providing a more complete and detailed history of the forest. Every property in the resulting dataset is represented by a list of values that represent the history of that property across all updates. For instance, if a point has the IFN species list [“Oak”, “Oak”, “Oak”, “Eucalyptus”], this represents that between 2010 and 2015 there was a change in the species present in that point. To gather this information, we intersected the points to evaluate with the auxiliary data sources, and only the precise point locations were taken into account.

Table 1 - Data sources used with their respective updates and the relevant properties that were extracted. (1*https://geocatalogo.icnf.pt/catalogo_tema3.html ; 2*https://geocatalogo.icnf.pt/catalogo_tema5.html ; 3*https://snig.dgterritorio.gov.pt/rndg/srv/por/catalog.search#/search?anysnig=COS&fast=index ; 4*https://land.copernicus.eu/pan-european/corine-land-cover)

Data Source Name	Updates	Relevant properties	Source
Inventário Florestal Nacional (IFN)	1995, 2005, 2010, 2015	id, land cover, species	1*
Burned Area Cartographic Data	Yearly from 1975 until 2018	id, year	2*
Carta de Uso e Ocupação do Solo (COS)	1995, 2007, 2010, 2015	id, land cover	3*
CORINE Land Cover (CLC)	1990, 2000, 2006, 2012	id, land cover	4*

2.3. Point Analysis Methodology

Next, we analysed each point individually supported by our own developed GEE application that, for a particular point, presents the auxiliary information, a Landsat satellite image of the area surrounding the point and the NDVI and NBR time series from 1985 onwards (**Figure 2)**. The index NDVI provides information about vegetation health, while NBR is used to confirm fire occurrences. The time series are interactive; one can click a particular observation and view the respective satellite image for obtaining additional visual context. The time series and RGB images are obtained from the Landsat program remote sensing imagery from missions 4, 5, 7 and 8, collection 1 tier 1 calibrated to top-of-atmosphere reflectance due to the adequate timespan for our objective (as detailed in **Figure 1**). While we could use TimeSync to create this type of dataset, GEE offers a more customizable product at the expense of requiring more user knowledge. The main difference was the temporal resolution of the analysed time series, TimeSync uses annual Landsat composites while our solution uses all the available observations.

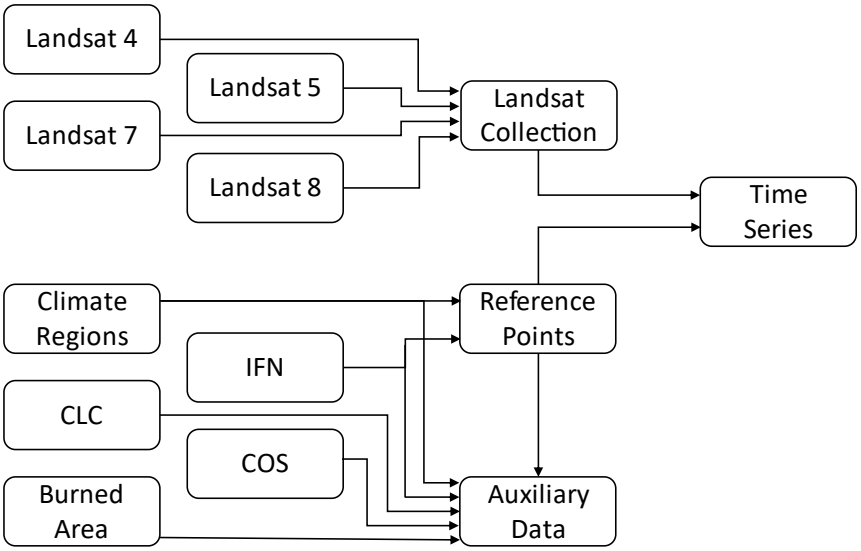


Figure 1 - Diagram of the workflow inputs

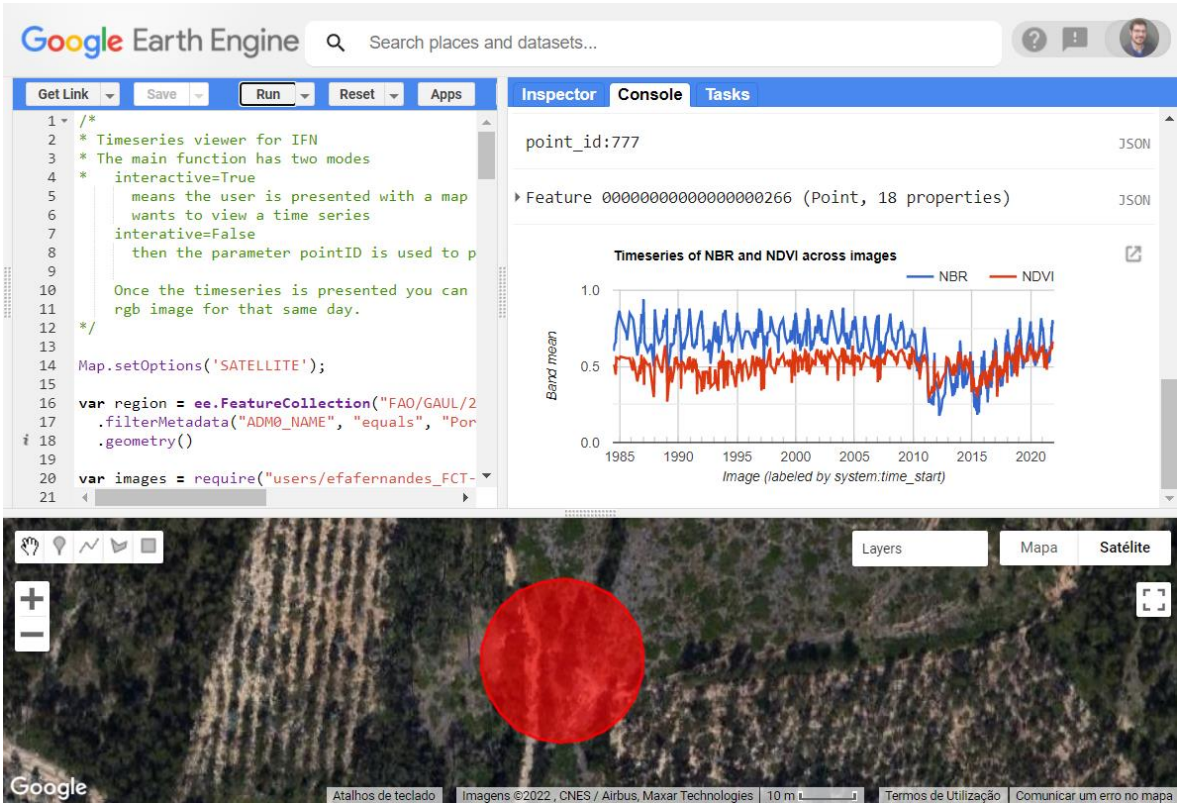
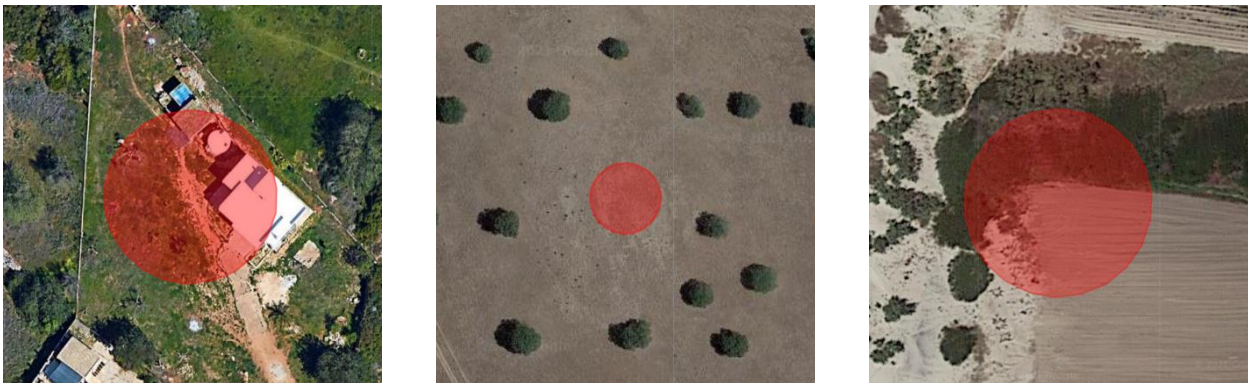


Figure 2 - GEE interface with interactive timeseries, satellite image and auxiliary information. The red circle indicates area to be analysed. In the time series plot we can observe two breakpoints.

The analysis of each point proceeded in three distinct steps described below: validation of the point, selecting changepoints and understanding the cause of the changepoint.

To validate one point, the surrounding circular area with a radius of 15m must meet four conditions. The area must not contain an artificial area, for example, if an area contains buildings or large roads as depicted in **Figure 3a** it is invalidated. The area must not be a low-density tree area, for example, if most of the area is bare soil or ground vegetation as shown in **Figure 3b** the point is considered invalid. The area must be a contiguous plot of land, if the area contains multiple plots of forest area with substantial differences as seen in **Figure 3c** then the

point is invalidated. The final condition is related to the time series, its analysis must be conclusive and decisive, if the time series is somehow erratic, hard to analyse or raises doubt, the point may be invalidated and classified as "other reason".



a. Artificial area ***b. Low-density area*** ***c. Different plots of land***

Figure 3 - Examples of three reasons for invalidation

After validating a point, the second step is to look at the time series and determine when the disturbances occur. In Figure 4, we find an NBR index time series with four clear moments where a disturbance occurs, these happen in the years a) 1987, b) 1997, c) 2005 and d) 2015. Not all time series are as clear as this one, some require a more thorough analysis.

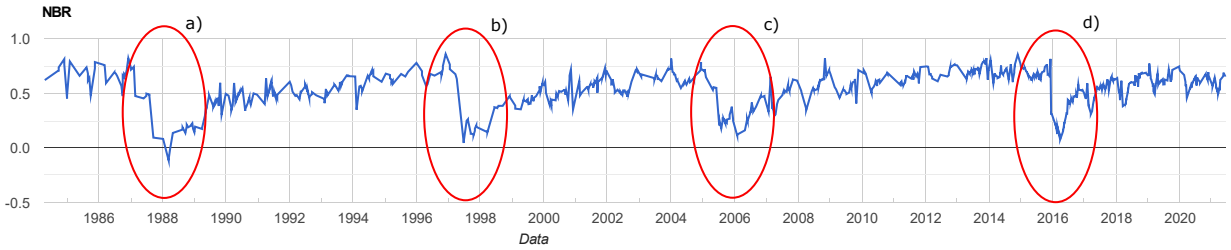


Figure 4 - Example of an NBR time series with four marked changepoints

The third step requires us to understand the cause of the disturbances, this is achieved by cross-referencing information throughout the auxiliary data described in **section 2.2** and using historic high-resolution satellite images from Google Earth Pro (GEP). For this dataset, four changepoint causes were considered: fire, harvest, unknown and other. Continuing with the example in **Figure 4** the auxiliary data says that fires occurred in the years 1987 and 2005, resulting in the annotation of Fire as the cause for changepoint a) and c). The remaining auxiliary data specifies that a eucalyptus forest has always occupied this point. Using GEP, it was possible to deduce that d) is most likely caused by a harvest. With the available data it was not possible to conclude what caused changepoint b) it will be marked as unknown.

3. Results

Having the dataset fully annotated we can turn our attention to its analysis. We will first examine the reason for invalidation, secondly their causes, thirdly we will assess the number of disturbances per valid point and finally evaluate the causes for the disturbances. It should be noted that oak, other softwoods, carob tree and acacias have an extremely small number of samples. Therefore, although represented in the figures, they will not be considered when extracting general conclusions based on species.

3.1. Number of invalidations

From a total of 998 sample points, 664 were considered valid and 334 were deemed invalid. This invalidation is not equally distributed through the species and consequently not uniformly distributed through the study area. When analysing **Figure 5**, we can see that the south and interior of the country have a higher density of invalidated points. The south of Portugal is dominated by cork oak, and we can see that this species is more

affected by invalidation than eucalyptus or maritime pine. The next subsection section will further explain this asymmetry.

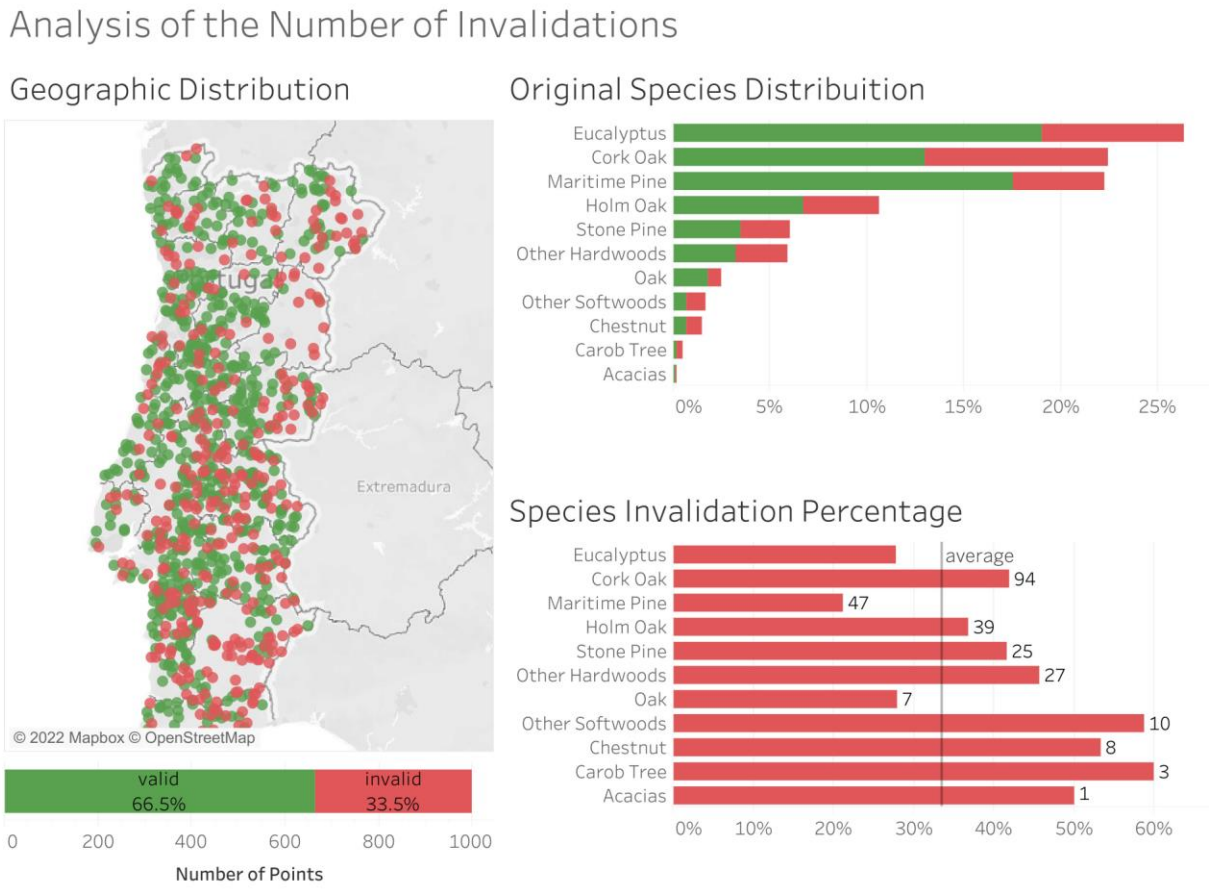


Figure 5 - Dashboard analysing the number of invalidations

3.2. Cause of invalidations

The primary cause for invalidation is low density closely followed by difficult analysis (see **Figure 6**). Once again there is a geographical asymmetry, with low density being predominant in the south and the interior of the country, this is the main reason these areas have a higher density of invalidated points, and consequently why cork oak has a relatively high number of invalidated points. These areas tend to be low density and populated by cork oak just as depicted in **Figure 3b**. We can see that the primary reasons species are invalidated are low density or other. Although this rule is hard to confirm or deny in species with small representation like acacias and carob trees. Additionally, eucalyptus has a high percentage of invalidations caused by complex plot limits.

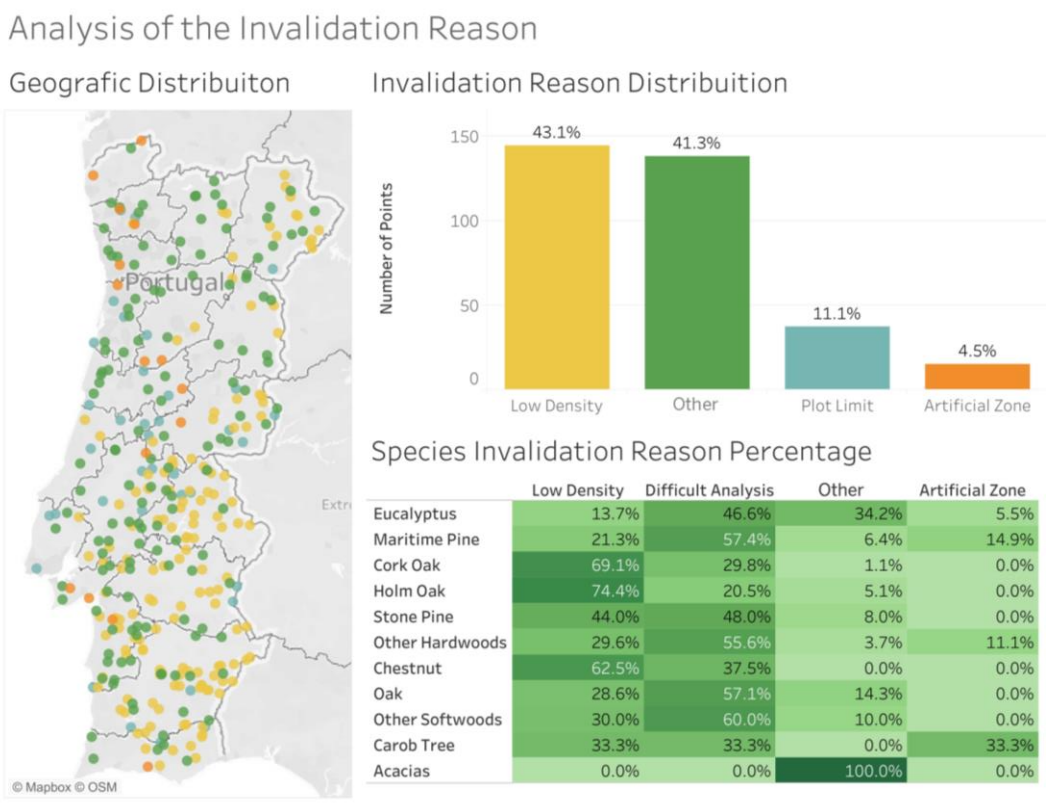


Figure 6 - Dashboard analysing the reason of the invalidation

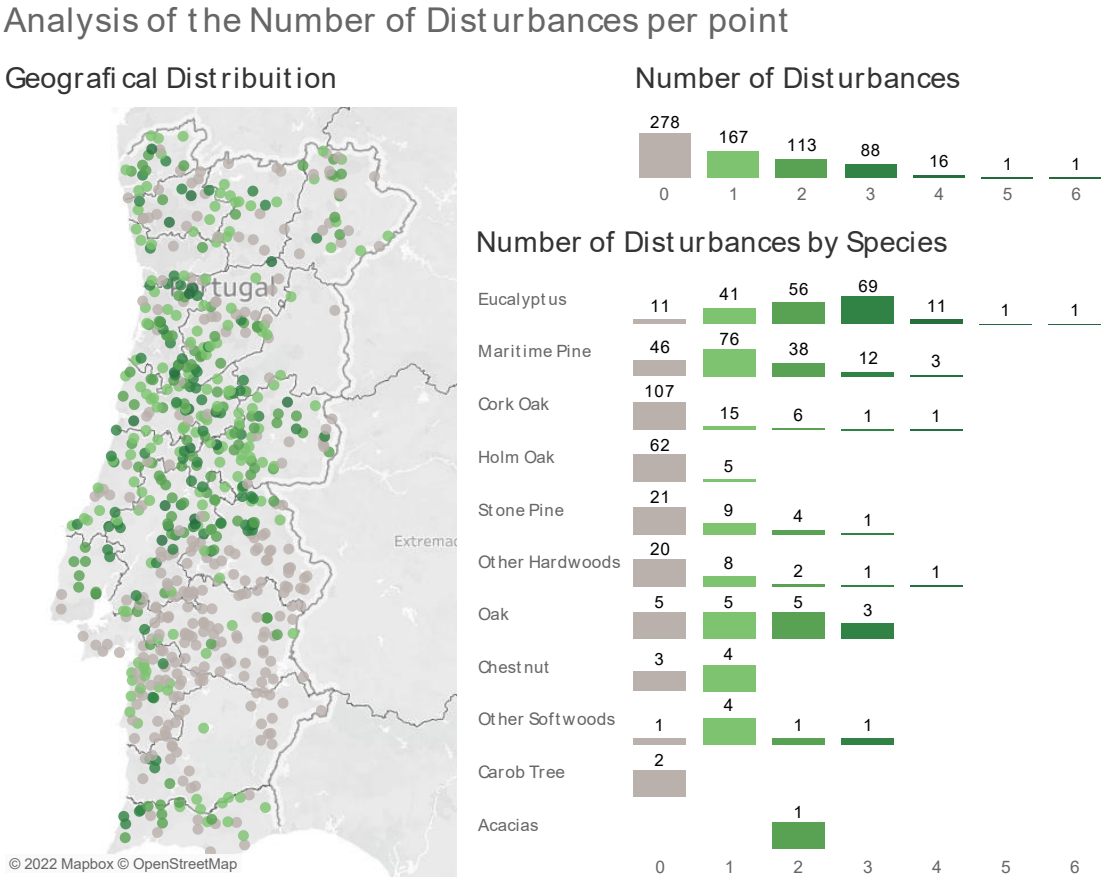


Figure 7 - Dashboard analysing the number of disturbances per point

3.3. Number of changepoints per point

The number of changepoints per point is summarized in **Figure 7,** showing that the mode of changepoints per point for the whole dataset is zero. However, the eucalyptus and the maritime pine classes have respectively a mode of 3 and 1 changepoints. This difference among species is reflected in the geographical distribution of the number of changepoints, the south has almost no changepoints while the center of the country is populated with points that have many disturbances, namely due to fire occurrences.

3.4. Cause of disturbance

Finally, the causes of the disturbances are reported in **Figure 8.** In our dataset, about 20% of the disturbances have an unknown origin, this makes it impossible to conclude if either fire or harvesting is the main cause of forest disturbances. Concerning the temporal distribution of causes, it can be concluded that unknown causes are predominant in the earlier years and tend to later disappear, while harvest cause has the exact opposite behavior. This is due to the lack of information about harvesting since the only way to validate this type of disturbance is either by a landcover change or by high-resolution satellite imagery. GEP only provides significant high-resolution photos after the year 2000. On the other hand, the fire records are in general complete throughout the years. For this reason, it can be assumed that many unknown causes may correspond to harvesting events.

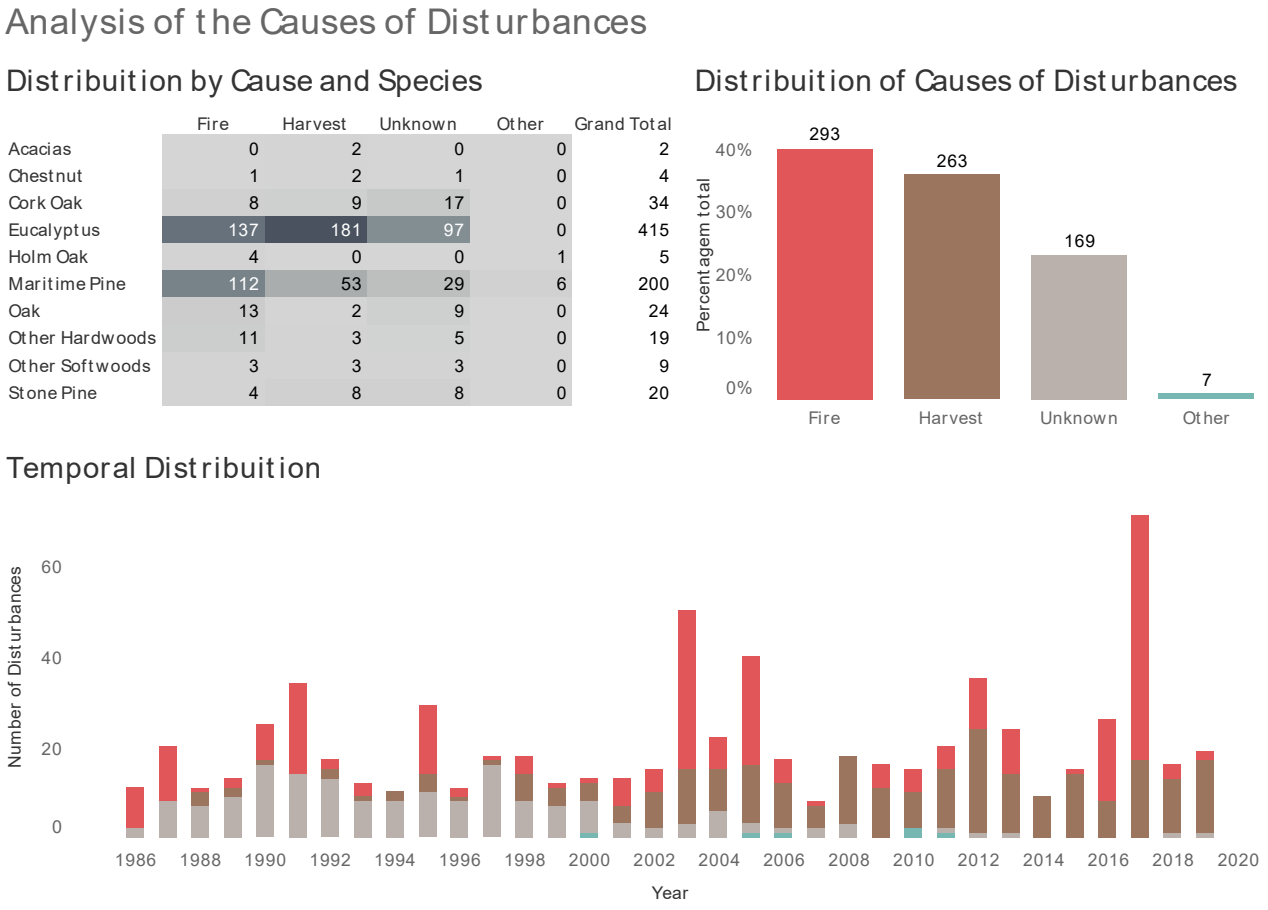


Figure 8 - Dashboard analysing the causes of disturbance

4. Discussion and Conclusions

In this paper, we described the creation of a curated dataset that contains 664 forest points generated from a stratified random sampling based on the tree species and climate zone that cover continental Portugal. Every point contains the auxiliary data used and a thorough list of all years where a disturbance occurs and the respective reason for the disturbance; the years span from 1986 until 2019.

While some of the breakpoints in the dataset can be obtained (with some work) from other sources, this dataset assures that if a breakpoint is not listed then almost certainly a disturbance has not occurred. Furthermore, due to the stratification, it is expected that the number and frequency and type of disturbances present are representative of real-world data for forest areas in Portugal. However, depending on the application, the number of geographic points may not be enough to provide full guarantees. Finally, this dataset only tracked severe disturbances, long term disturbances like droughts were not considered, and thus it may not meet the requirements of some applications.

This dataset can be used in several applications. First, and the main reason motivating the creation of the dataset, it can be used to evaluate algorithms that detect forest disturbances for example or a more generical use as a dataset for evaluating changepoint detection in time series. Finally, it may be employed for differentiating types of forest recovery.

5. References

Brooks, E. B., Wynne, R. H., Thomas, V. A., Blinn, C. E., & Coulston, J. W. (2014). On-the-Fly Massively Multitemporal Change Detection Using Statistical Quality Control Charts and Landsat Data. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(6), 3316–3332. https://doi.org/10.1109/TGRS.2013.2272545

Cohen, W. B., Healey, S. P., Yang, Z., Stehman, S. V., Brewer, C. K., Brooks, E. B., Gorelick, N., Huang, C., Hughes, M. J., Kennedy, R. E., Loveland, T. R., Moisen, G. G., Schroeder, T. A., Vogelmann, J. E., Woodcock, C. E., Yang, L., & Zhu, Z. (2017). How similar are forest disturbance maps derived from different Landsat time series algorithms? *Forests. 8: 98.*, *8*, 98. https://doi.org/10.3390/f8040098

Cohen, W. B., Yang, Z., & Kennedy, R. (2010). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync — Tools for calibration and validation. *Remote Sensing of Environment*, *114*(12), 2911–2924. https://doi.org/10.1016/j.rse.2010.07.010

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Huang, C., Goward, S. N., Masek, J. G., Thomas, N., Zhu, Z., & Vogelmann, J. E. (2010). An automated approach for reconstructing recent forest disturbance history using dense Landsat time series stacks. *Remote Sensing of Environment*, *114*(1), 183–198. https://doi.org/10.1016/j.rse.2009.08.017

Kennedy, R. E., Yang, Z., & Cohen, W. B. (2010). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — Temporal segmentation algorithms. *Remote Sensing of Environment*, *114*(12), 2897–2910. https://doi.org/10.1016/j.rse.2010.07.008

Vogelmann, J. E., Xian, G., Homer, C., & Tolk, B. (2012). Monitoring gradual ecosystem change using Landsat time series analyses: Case studies in selected forest and rangeland ecosystems. *Remote Sensing of Environment*, *122*, 92–105. https://doi.org/10.1016/j.rse.2011.06.027

Zhu, Z., & Woodcock, C. E. (2014). Continuous change detection and classification of land cover using all available Landsat data. *Remote Sensing of Environment*, *144*, 152–171. https://doi.org/10.1016/j.rse.2014.01.011