**Short Term Scientific Mission (STSM) on improving Pinus pinea L. cone yield modeling for Portugal**

COST Action: FP1203–European Non Wood Forest Products (www.nwfps.eu)

Dates of the mission: 18/05/2015 to 28/05/2015

Fellow: João Pedro Abranches Freire

Sending organization and supervisor: University of Lisbon, Instituto Superior de Agronomia, Maria Margarida Branco de Brito Tavares Tomé

Host organization and supervisor: National Institute for Agricultural and Food Research and Technology (INIA), Rafael Calama Sainz and Sven Mutke Regneri

Index

# 1 Background and Purpose

*Pinus pinea* L. is a species of growing economic importance both in Portugal and in Spain because of its main product, pine cones.

Masting, the intermittent and pulsed set of seeds, is a common and important phenomenon in many plant species, like *Pinus pinea* L.

There is a diverse range of factors that affects masting events that is necessary to understand and consider into cone modeling.

The exchange of knowledge within the Mediterranean countries, where the pine nut industry has a great economic value, is very important to identify those factors and to use them into cone modeling.

This Short Term Scientific Mission (STSM) had as main purposes:

- Model cone yield considering the effect of climate on masting events;
- Learn recent methodologies to calibrate mixed models;
- Strengthen the cooperation with the modeling team of the Forest Research Centre in the National Institute for Agricultural and Food Research and Technology of Spain (INIA).

# 2 Description of the work carried out during the STSM

## 2.1 Material

### 2.1.1 Permanent plots for cone production

In 2004 and 2005, 73 permanent plots were established into the most productive region for cone production in Portugal, the V Portuguese Provenance Region – “Charneca Miocénica e Pliocénica dos Vales do Tejo e do Sado”. (Carneiro et al., 1998).

The trees from the plots have been measured in 2004 or 2005, 2011 and 2015. Diameter at breast height, total and crown base height and crown diameter were measured at all trees of the plots independently of the specie.

Cones were harvested mostly in 2004/05 till 2008/09 and in 2013/14 and 2014/15 campaigns. In the rest of the campaigns cones have only been collected at few trees (table 1). There are more cone yield data from properties CASC and M but they could not be used because they were still not available at the time this work was done.

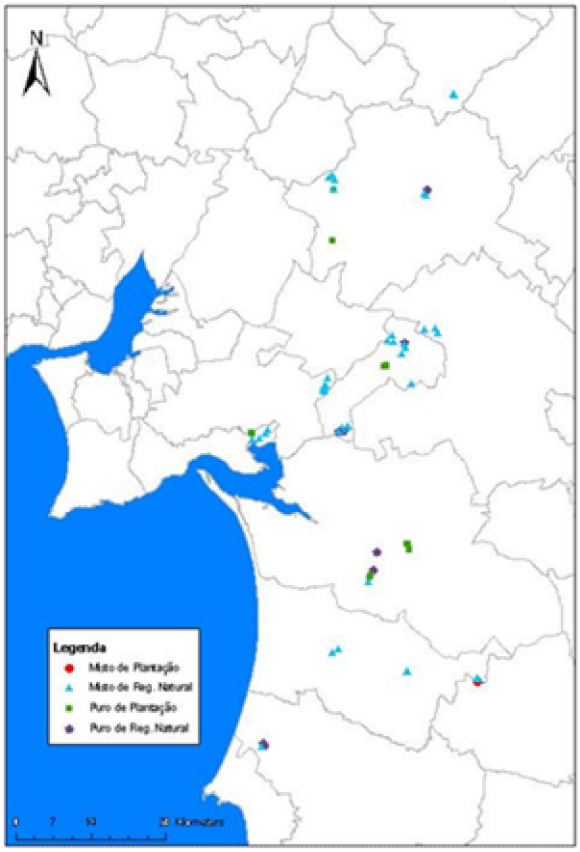



**Figure 1.** Distribution of plots in the study area

During the process of analyzing data we were able to detect two bumper crops in the plots at least 20 km from the coast. They have occurred in 2004/05 and in 2010/11 campaigns. The last was the biggest bumper crop ever recorded, because of that it is very difficult to model. In plots near the sea a bumper crop could be detected in the 2013/14 campaign.

Like referred in Calama et al. (2011), we have verified a huge influence of the bumper crops over the following three crops, mainly over the third one with the depletion of cone probably due: 1) to the inhibition of flower induction in the bumper crop year as a consequence of the lack of nutrients that may be shifted to the growing cones and/or 2) the climatic conditions that leaded to the bumper crop have no influence over this campaign. We have verified that the greater is the bumper crop the lower is the production of cones three years after it comparing to the average production of the site.

**Table 1.** Trees in which cones were collected

| Property ID | Campaigns | | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2004/05 | 2005/06 | 2006/07 | 2007/08 | 2008/09 | 2009/10 | 2010/11 | 2011/12 | 2012/13 | 2013/14 | 2014/15 | |
| A | | 16 | 5 | | | | | | | | 15 | 36 |
| BS | | 64 | 39 | | | | | | | | | 103 |
| CASC | | 89 | 89 | 82 | 53 | 49 | 46 | 22 | 48 | | | 478 |
| CB | 124 | 98 | 111 | 22 | | | | | | 12 | 140 | 507 |
| CC | | | 65 | | | | | | | | | 65 |
| G | | | 83 | | | | | | | | | 83 |
| HE | 45 | | 107 | 35 | | | | | | 90 | 88 | 365 |
| M | | 23 | 23 | 19 | 23 | 23 | 23 | 21 | | | 22 | 177 |
| MBSA | 57 | 257 | 170 | | | | | | | | 129 | 613 |
| MM | | 79 | | | | | | | | 179 | 179 | 437 |
| MNV | 227 | | | | | | | | | | | 227 |
| OM | | 143 | 143 | | | | | | | | | 286 |
| PS | 8 | 29 | 56 | 13 | | | | | | | | 106 |
| QS | 29 | 85 | 104 | 16 | | | | | | | 41 | 275 |
| VM | | 93 | 106 | | | | | | | | 43 | 242 |
| VOB | | 330 | 330 | 34 | | | | | | | | 694 |
| Total | 490 | 1306 | 1431 | 221 | 76 | 72 | 69 | 43 | 48 | 281 | 657 | 4694 |

### 2.1.2 Weather variables

Since there are many missing data from meteorological stations, we have considered interpolated data produced by Haylock et al. (2008), version 11.0, considering a resolution of 0.25º kindly made available by Climate Change Impacts Adaptation & Modeling Research group (CCIAM) from Faculdade de Ciências at Lisbon University.

The nearest points to the permanent plots were selected from the referred grid totalizing 8 points.

For each point, the influence of the limiting weather variables precipitation and maximum temperature on cone growth was studied.

## 2.2 Methods

### 2.2.1 Response variable

In each campaign we have considered individual cone yield per tree expressed by the total number of cones (*nc*) and their fresh weight (*wc*). To model cone yield we have considered two different approaches:

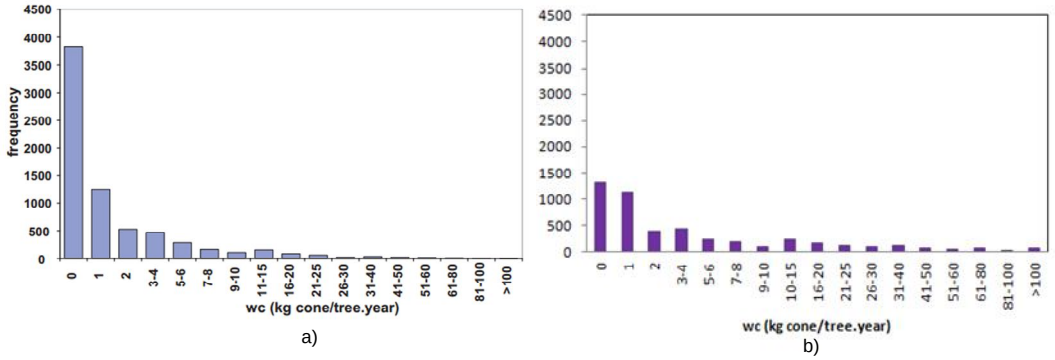
- Since *nc* and *wc* are ruled by distinct climate variables we have opt to model them separately considering the estimation of *nc* over the modulation of *wc* and

– In order to compare this work with the carried out by Calama et al. (2011), Freire (2009) and Rodriges et al. (2014) we have considered the response variable *wc* without taking into account the estimation of *nc*.

The model presenting the better response was selected.

Like in Calama et al. (2011) the distribution of frequencies for both *wc* and *nc* variables did not fulfill the standard normality assumption, displaying:

– Asymmetry: empirical distribution is significantly skewed towards the higher values of the variable, with a massive number of observations showing smaller values of cone production, and only a small number of trees in a few years giving very large crops and forming a long tail to the right.
– Zero inflation: the distribution displays a strong mode at zero (corresponding to null production by sampled trees), comprising 28.12% of the observations in the fitting data set, against the 54.49% at Calama et al. (2011) (figure 2).
– Truncation: given the nature of the response variable, negative values are not possible.



a) b)

**Figure 2.** Histogram of observed frequencies of annual cone production of single trees (kg tree−1 year−1) including all the observations from the three data sets at Spanish a) and Portuguese b) level.

Furthermore, the hierarchical structure of the data (repeated observations from trees nested in sample plots within Natural Units) implies a lack of independence among observations, which prevented us from using estimation methods based on ordinary least squares minimization.

### 2.2.2 Modelling approach

The data with excess zeros are very common within the ecology and forest science. This type of data presents a series of particularities which prevent the application of classical statistical techniques based on the assumption of normality, as in the case of the least square regression (Calama et al., 2012).

In the analysis and development of models when the response variable has a large number of zeros, in the first place it is necessary to identify the theoretical distribution model to which resemble the data, by establishing a linear relationship through a link function between some of the

parameters characterizing the distribution, usually the mean, and one or several explanatory variables, categorical or continuous, through a series of parameters that are estimated by maximum likelihood methods (Affleck, 2006, Calama et al., 2012). In this sense, the resolution of such models can be considered a case of generalized linear models (McCullagh and Nelder, 1989; Tu, 2002, Calama et al., 2012).

To select the distribution that best fits the data we have followed the methodology proposed by Calama et al. (2012) taking into account all data. For the modeling of the cone number the distributions Poisson, Negative Binomial (NB), zero inflated Poison and Negative Binomial (ZIP and ZINB respectively) were tested. For modeling cone weight the distributions log-normal and zero inflated log-normal (LN and ZILN respectively) were tested.

Selected the distribution, we tested the introduction of variables in order to maximize the logarithm of the likelihood function.

### 2.2.3 Covariate selection

The explanatory covariates may or may not be common to both the occurrence and intensity models. In modeling stone pine cone production, we evaluated different groups of possible explanatory variables:

- Tree size: basal area at breast height (*g*) and diameter at breast height (*d*), total height (*h*) and crown ratio (*cr*), crown width (*cw*)
- Stand attributes: number of stems per ha (*N*), basal area (*G*), Stand Density Index (*SDI*), quadratic mean diameter (*dg*), dominant height (*hdom*) and mean distance between trees (*Mdist*).
- Distance-independent competition indices: basal area of trees larger than the subject tree (*BAL*), *d*/*dg*, *g*/*G*

The above variables are assumed to account for spatial variability in cone production and were considered constant since 2011 because measurements at tree level are taking place. Temporal variability in cone production was explained by evaluating different weather attributes over the course of the study period: monthly rainfall, average, maximum and minimum monthly temperatures. Given the three-year duration of cone development, the weather variables up to four years prior to cone maturation were evaluated.

To select the weather attributes to test first we eliminated the influence of site by dividing the cone production at each year by the mean cone production, producing the variable *ratio number of cones* (RNC) that reflects the relative masting pattern among years. Afterwards graphically we have compared RNC over time with climate variables.

### 2.2.4 Evaluation of the model

Modeling was performed using the SAS system and at this first approach the tested models were ordered by statistic Akaike's Information Criterion (*AIC*).

The 10 models with lowest *AIC* with all variables with significant and meaningful parameters were analyzed taking into account AIC, the adjusted $R^2$ $\left(R^2_{Aj}\right)$, the mean of the differences between mean observed NTP and mean estimated NTP at each campaign $\left(\overline{DifCl}\right)$ and mean estimate of the absolute differences between observed NTP and mean estimated NTP at each campaign ($|DifCl|$).

$$R^2 = 1 - \frac{SQ_{res}}{SQ_{tot}} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} \quad (1)$$

$$R^2_{Aj} = 1 - \frac{n-1}{n-(k+1)}(1 - R^2) \quad (2)$$

$$\left|\overline{DifCl}\right| = \left| \frac{\sum_{j=1}^{ncl}\left(\frac{\sum_{k=1}^{nycl} y_{kj}}{nycl_j} - \frac{\sum_{ki=1}^{nycl} \hat{y}_{kj}}{nycl_j}\right)}{ncl} \right| \quad (3)$$

$$\overline{|DifCl|} = \frac{\sum_{j=1}^{ncl}\left(\left|\frac{\sum_{k=1}^{nycl} y_{kj}}{nycl_j} - \frac{\sum_{ki=1}^{nycl} \hat{y}_{kj}}{nycl_j}\right|\right)}{ncl} \quad (4)$$

where *n* represents the total number of observations, $y_i$ and $\hat{y}_i$ are the observed and predicted values of number of for observation *i,* $\bar{y}$ is the mean value for the response variable, $y_{kj}$ e $\hat{y}_{kj}$ are the observed and predicted values of number of for observation *k* at campaign *j*, $nycl_j$ and $ncl$ are the total number of observations *j* at campaign *k* and the total number of campaigns.

Since it was taken into account several statistics it was considered a multi-criteria analysis to select the best model.

To enable the treatment of the statistics first it was necessary to normalize each value of each statistic by subtracting to each statistic of each model the mean of the statistic and dividing by the standard deviation (s).

$$NAIC_j = \frac{AIC_j - \overline{AIC}}{\sigma AIC};\ NR^2_{Ajj} = \frac{R^2_{Ajj} - \overline{R^2_{Aj}}}{\sigma R^2_{Aj}};\ N\left|\overline{DifCl_j}\right| = \frac{\left|\overline{DifCl_j}\right| - \overline{\left|\overline{DifCl}\right|}}{\sigma\left|\overline{DifCl}\right|};\ N\overline{\left|DifCl_j\right|} = \frac{\overline{\left|DifCl_j\right|} - \overline{\left|\overline{DifCl}\right|}}{\sigma\left|\overline{DifCl}\right|} \quad (5)$$

It was selected the highest $NR^2_{Ajj}$ (Max$NR^2_{Aj}$) and the lowest $NAIC_j$ (Min$NAIC$), $N\overline{DifCl_j}$ (Min$\overline{NDifCl}$ and $N\overline{\left|\overline{DifCl_j}\right|}$ (Min$N\overline{|DifCl|}$)

For each model *j* was calculated the statistic MCS given by the expression (6):

$$MCS_j = \text{Max}NR^2_{Aj} - NR^2_{Ajj} + 2\left(NAIC_j - \text{Min}NAIC\right) + 3\left(N\left|\overline{DifCl_j}\right| - \text{Min}\left|\overline{NDifCl}\right|\right) + \quad (6)$$

$$+4\left(N\overline{\left|\overline{DifCl_j}\right|} - \text{Min}N\overline{|DifCl|}\right)$$

Into equation 6 we have attributed increasing importance to $R^2_{Ajj}$, AICj, $\left|\overline{DifCl_j}\right|$ and $\overline{\left|DifCl_j\right|}$.

## 2.3 Work performed during the STSM

### 2.3.1 First week

During the first week of the Short Term Scientific Mission several climate variables were calculated or interpolated that could influence the fluctuation of cone number over time and those that had a performance similar to RNC were selected graphically.

This work was performed under the supervision of Dr. Sven Mutke.

### 2.3.2 Second week

At the second week the performance of the listed distributions to the data was studied and selected the one with higher performance to model the number of cones following the work developed by Calama et al. (2011).

It was tested the covariates listed above and the climate variables selected into the first week into the selected distribution.

This work was performed under the supervision of Dr. Rafael Calama.

# 3 Description of the main results obtained

### 3.1.1 First week

49 climate variables were tested since four years before harvest till two years before harvest to all plots with at least three cone collections. In table 2 the selected variables are shown.

It can be seen that the selected variates were the rainfall in November one year before primordia formation and the temperature at the year of flower emergence or at the summer after flowering that takes place in Portugal during the period April – May.

The greater the precipitation, the higher the production of cone, and the higher the maximum temperature or maximum number of days with temperatures greater than 30 or 35 degrees, the lower the amount of cone.

We can postulate that the precipitation four years before harvesting influences the number of cones that are differentiated and the maximum temperature two years before harvest influences the survival rate of the flowers.

**Table 2.** Climate variable selected to model the number of cones

| Variable | Description |
|---|---|
| pp_nov_4 | Precipitation in November four years before harvest |
| MeanMaxT_Jul_Sept_2 | Mean of maximum temperatures between July and September two years before harvest |
| NDMaxT_Jul_Sept_2_M30 | Number of days between July and September with maximum temperature higher than 30ºC two years before harvest |
| NDMaxT_Jul_Sept_2_M35 | Number of days between July and September with maximum temperature higher than 35ºC two years before harvest |
| NDMaxT_Year_2_M30 | Number of days in the year with maximum temperature higher than 30ºC two years before harvest |
| NDMaxT_Year_2_M35 | Number of days in the year with maximum temperature higher than 35ºC two years before harvest |

### 3.1.2 Second week

We have tested the distributions mentioned in point 2.2.2 to model cone number and cone weight without considering covariates. The results are presented into table 3.

It can be seen that the distribution with better performance (with lower AIC) was in both cases the Binomial Negative (NB), followed by the Zero Inflated Binomial Negative (ZINB) and by the Hurdle ZINB. Since the simplest distribution is also the one with better performance we have selected it although the other two had performances quite similar.

The other distributions have had a performance quite lower.

**Table 3.** Performances of the distributions without considering covariates

| Distribution | Number of cones | | Weight of cones | |
|---|---|---|---|---|
| | -2log | AIC | -2log | AIC |
| Poisson | 353361 | 353363 | 117398 | 117400 |
| Binomial negative (NB) | 26699 | 26703 | 19760 | 19764 |
| Zero inflated Poisson (ZIP) | 332785 | 332789 | 112800 | 112804 |
| Hurdle ZIP | 332785 | 332789 | 112800 | 112804 |
| Zero inflated NB (ZINB) | 26699 | 26705 | 19760 | 19766 |
| Hurdle ZINB | 27223 | 27229 | 20517 | 20523 |
| Zero Inflated Log-normal (ZILN) | | | 28747 | 28753 |

After testing different sets of covariates into NB distribution, considering the methodology proposed at point 2.2.3, we were able to select the model present at equation (7).

$$N = e^{c0+c1*d+c2*h+c3*Mdist+c4*BAL+c6*NDMaxT\_Jul\_Sept\_2\_M30+c5*pp\_nov\_4} \quad (7)$$

At table 4 are shown the statistics of the selected model.

Table 4. Statistics of the selected model

| Statistic | Value |
|---|---|
| -2 Log Likelihood | 23922 |
| AIC (smaller is better) | 23940 |
| AICC (smaller is better) | 23940 |
| BIC (smaller is better) | 23987 |
| $NR^2_{Aj}$ | 0.6241 |
| $\overline{\lvert \mathrm{DifCl} \rvert}$ | 0.1273 |
| $\overline{\lvert \mathrm{DifCl} \rvert}$ | 22.2969 |

Since we have considered this model as exploratory one we have opt to not consider at this work the value of the parameters.

At figure 3 it can be seen the mean number of cones and the mean predicted number of cones at each campaign. The model could predict well the number of cones except at the bumper crop of 2010/2011 and at the campaign 2012/2013 that is directly related to the other.

It is important to take into account that the campaign 2013/14 was a bumper crop at the plots near the sea and we have several plots in that zone. This can explain the increase into cone production from 2012/13 to 2013/14 at figure 3.

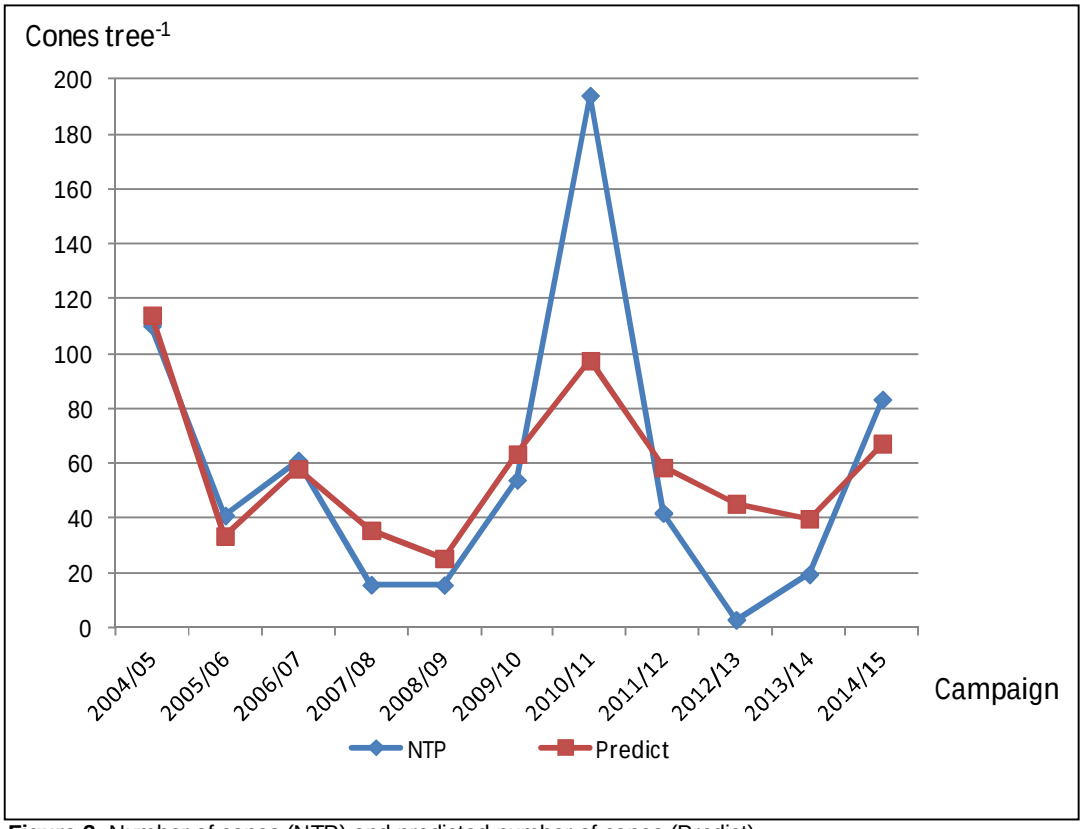



**Figure 3.** Number of cones (NTP) and predicted number of cones (Predict)

## 4 Future collaboration with host institution (if applicable)

From figure 3 it can be seen that the selected model did not estimate well cone number of the campaigns 2010/11 and 2012/13. This was a reality for all models tested. For this reason we have considered that is important to test the influence of new combinations of weather variables over cone production. We are going also to simulate the occurrence of bumper crops since there it influence greatly cone productions for three consecutive.

## 5 Foreseen publications/articles resulting or to result from the STSM

As result of this STSM at least one publication will be realized modelling cone production.

## References


Affleck, D.L.R., 2006. Poisson mixture models for regression analysis of stand level mortality. Can. J. For. Res. 36, 2994-3006.

Calama, R., Manso, R., Tomé, J., 2012. Cómo modelizar datos com exceso de ceros? Métodos y aplicación a la investigación forestal. Cuad. Soc. Esp. For. 34, 55–65.

Calama, R., Mutke, S., Tomé , J., Gordo, J., Montero, G., Tomé M., 2011. Modelling spatial and temporal variability in a zero-inflated variable: The case of stone pine (Pinus pinea L.) cone production. Ecological Modeling 222, 606-618.

Carneiro, M., d'Alpuim, M., Rocha, M. 1998. Delimitação e caracterização de regiões de proveniência de *Pinus pinea* L.em Portugal. Silva Lusitana 6, 129–160.

Freire, J. 2009. Modelação do crescimento e da produção de pinha no pinheiro manso. PhD. Thesis. ISA/UTL.

Haylock, M.R., Hofstra, N., Klein Tank, A.M.G., Klok, E.J., Jones , P.D., New, M., 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. Journal of Geophysical Reseach, 113, D20119.

McCullagh, P., Nelder, J., 1989. Generalized linear models. Second Edition. Chapman and Hall/CRC. Boca Ratón.

Rodrigues A., Silva G., Casquilho M., Freire, J., Carrasquinho, I., Tomé M., 2014. Linear mixed modelling of cone production in stone pine in Portugal. Silva Lusitana, 22 (1), 1–27;

Tu, W., 2002. Zero-inflated data. In: El-AH. Shaarawi & W.W. Piegorsch (eds.) Encyclopedia of Environmetrics: 2387-2391. John Wiley and Sons. Chichester.

## Annex 1: Confirmation by the host institution of the successful execution of the STSM

MINISTERIO
DE ECONOMÍA
Y COMPETITIVIDAD

INIA
Instituto Nacional de Investigación
y Tecnología Agraria y Alimentaria
Subdirección General de Investigación y Tecnología
**Centro de Investigación Forestal**

# CERTIFICATE

Dr. Rafael CALAMA, Senior Researcher, and Dr. Sven MUTKE, Head of Service for Forest Industries, in the Forest Research Centre INIA-CIFOR at the National Institute for Agricultural and Food Research and Technology, Madrid,

CERTIFICY:

That Dr. João FREIRE from University of Lisbon - Institute Superior of Agronomy, Portugal stayed two weeks from May 18th to May 28th 2014 for a short-term scientific mission (STSM) under the COST Action FP1203 "European non-wood forest products (NWFPs) network" in our Centre, collaborating with us about the topic *Improvement of* Pinus pinea *L. cone yield modelling for Portugal* (COST STSM Reference Number: COST-STSM-FP1203-25345).

The stay concluded with success, achieving an advance in data analysis and model building based on Portuguese cone yield and climate data and Spanish model architecture and methods. A first layout of a common publication was plotted during these weeks, discussing common patterns and differences of weather influence on cone yield in both countries. The topic of the STSM is of interest for both institutions, also with a view to other possible future collaborations in the framework of the WG2 and TF2 of the COST Action FP1203.

In Madrid, May 28th 2015

SUBDIRECCION GENERAL DE INVESTIGACION Y TECNOLOGIA
CENTRO DE INVESTIGACION FORESTAL
CIFOR-INIA

Fdo.: Dr. Rafael CALAMA

Fdo: Dr. Sven MUTKE

Dr. Sven Mutke Regneri
Jefe de Servicio
Centro de Investigación Forestal CIFOR
email mutke@inia.es

INIA CIFOR
Ctra. de La Coruña, km 7,5
28040 MADRID
TEL: 91 347 6874
FAX: 91 347 6767