A Model for Predicting Yield of Seeds in Ficus Fruits

H.L.D.K. Jayarathna and L.S. Nawarathna*
Department of Statistics and Computer Science
University of Peradeniya
Peradeniya, Sri Lanka
*lakshikas@pdn.ac.lk

W. A. I. P. Karunaratne
Department of Zoology
University of Peradeniya
Peradeniya, Sri Lanka

***Abstract*— *Ficus* is one of the largest plant genus which has an ecological significance due to the presence of "keystone" species. Availability of its sole mutualistic wasp pollinator and the effect of non-pollinator wasps determine the availability of seeds in *Ficus* fruits to produce the next generation of each species. In most of the previous studies on the yield of seeds of *Ficus* fruits, seeds have been counted manually, which is a time consuming and hectic process. Therefore, the main objective of this study is to introduce a model for predicting the yield of seeds in two *Ficus* species. Local polynomial regression, generalized additive models, and Poisson regression models were used for constructing these models to predict the number of seeds per fruit. Two generalized additive models which were constructed for Kandy municipal & Thumpane were best described with lower mean square error values of testing samples and moderately large R^2 values when Fruit length was taken as a single predictor for both models. Poisson regression model gives a better result for modeling in *Ficus callosa* with min-max scaled variables, with a lowest mean square error value of the testing sample. Models which were built up for areas give the best prediction values when the yield of seeds less than 1000. By using local polynomial regression curves, it was identified that both biasedness and variance can be optimized using optimal bandwidth which was calculated by plug in the rule and it gives freedom to the flow of data by keeping non-parametric qualities.**

Keywords; nonparametric regression; bandwidth; pollinator wasps; smoothing functions.

I. Introduction

Ficus is one of the large plant genus and it has an environmental, evolutionary and conservation interest. About 10% of all bird species and 6% of all mammal species are eating *Ficus* fruits and most of them are capable of dispersing *Ficus* seeds [1]. The initial stage of *Ficus* fruits are commonly known as fig and its botanical name is *synconium*. Fig is a urn-like structure which is lined by tiny flowers acts firstly as inflorescence and following pollination seeds develop into fruits. All *gynodioecious Ficus* species are Old World species [2]. Monoecious figs have both male and female flowers while *gynodioecious* has two types of trees to produce flowers [3]. There is a unique identification for *Ficus* species as a "keystone" species. *Ficus* species are best known because of the relationship with pollinating wasps (*Hymonoptera: Agaonidae*) [1]. Because of this pollinating process two types of seeds namely damaged seeds (galls) and undamaged seeds(seeds) result. Therefore, seeds and galls play a major role in studies of *Ficus* trees. But the problem is with a large number of the yield of seeds; counting seeds is time consuming and hectic process [4]. As a solution for this issue, a statistical approach has been used for predicting the yield of seeds. The study has been carried out covering three sampling areas which are Kandy municipal area, University of Peradeniya and Thumpane. Two types of species namely *Ficus callosa* and *Ficus racemosa* were subjected to this study. It can be identified that all three areas have approximately same climatic features which help for controlling other factors cause to yield of seeds. Same aged trees were selected for controlling phytomorphological factors. Nonparametric statistical methods have been introduced to predict the yield of seeds using easily readable dimensional measurements. Since the yield of seeds is a count, Poisson Regression is also performed in this study [5].

II. Materials and methods

Using the simple random sampling method samples were collected with excess since there can be damages during transportation and insect bites. Recorded data were split into two samples as testing and training in the ratio 3:5. A statistical summary was obtained for getting a clear idea about data set and outliers were identified using R statistical software packages. Since three dimensional values are recorded as fruit length, fruit diameter1 and fruit diameter2 for each fruit, multicollinearity was checked using Variance Inflation Factor (VIF) value. For further studies, all dimensional variables were transformed into values between 0 and 1 using min-max transformation and those variables were defined as scal.Fl, scal.Fd1 and scal.Fd2 respectively. Since the yield of seeds is a count, Poisson Regression was applied by taking log link function. Poisson models give a structure combining with generalized linear models such that,

$$\ln(\text{Total seeds}) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \qquad (1)$$

where X_1 = Fruit length, X_2 = Fruit diameter 1 and X_3 = Fruit diameter 2 which were scaled between (0, 1).

In (1), β_j's represented coefficients of the three independent variables X_j's. Nonparametric regression approach [6] is used since the violation of normality by all dimensional measurements and seed amount. With modifications of running line smoothers kernel regression has introduced a new method for taking the size of the neighborhood called bandwidth instead of the equally weighted neighborhood for constructing curves. With a kernel smoother, X_i ‘s weights are dependent on their distance from X_0 particular data point. The equation assigning weight for X_i for predicting value at X_0 is,

$$W_{0i} = \left(\frac{C_0}{\lambda}\right) K\left(\frac{|X_0 - X_i|}{\lambda}\right) \qquad (2)$$

In (2), K(t) is an even function decreasing in t, λ is the bandwidth and C_0 is a constant that makes weight sum to 1.The general form of kernel smoother is written as,

$$\widehat{Y_J} = \sum_{i=1}^{n} W_{ij} \, Y_j \qquad (3)$$

In (3), $\widehat{Y_J}$ represents estimated value and W_{ij} and Y_j define weight and actual variable value respectively. Local polynomial regression [7] was identified as combination of kernel regression and running line smoother. In this method overlapping neighborhoods are used. An estimated value can be calculated using Nadaraya-Watson estimation based on the above results. Also, the importance of bandwidth was checked by adjusting it with optimal value and effect for variance and biasedness were studied. Curves for adjusted optimal bandwidth were drawn at same graph for clear identification. Instead of local polynomial regression, generalized additive model (GAM) was also studied as likelihood-based regression model [8, 9]. While $\sum \beta_j X_j$ is used in linear form, a smoothing function was introduced as $\sum S_j$ (X_j). In generalized additive models, “s” represents a smoothing function and is not a parameter. Smoothing function is built up with support the of basic splines after a complex computational process. Degree of smoothing of smoothing function is calculated by penalized basic splines by introducing a parameter called “λ” which is known as a smoothing parameter. These functions are estimated using scatterplot smoothers in an iterative procedure called local scoring algorithm. This method uncovers nonlinear covariate effects. Degree of smoothing is calculated using basic splines and estimated degree of freedom is given with the smoothing function [8]. Model validation was done using testing samples of 75 observations. Akaike’s Information Criteria (AIC) and Bayesian Information Criteria (BIC) were used for validating the seed predicting generalized additive models [9, 10].

III. Results

It was identified that each dimensional variables, seeds, and galls are positively or negatively skewed using dispersion measurements. Anderson Darling test resulted in a p-value less than α ($p<0.05$) which implies a violation of normality. Therefore, Local polynomial regression and GAMs were used for finding a nonparametric solution. Since the yield of seeds is a count, Poisson Regression was also applied. The model which is predicting the yield of seeds for *Ficus callosa;* with scaled variables in fruit length, fruit diameter1, and fruit diameter2 which values are lying between (0,1) gave the least mean squared error(MSE) value of the test sample.

TABLE I POISSON MODELS FOR SPECIES

SPECIES	**MODEL**	**MSE**	**AIC**	**BIC**
Ficus racemoca	ln(Total seeds) = 7.411115 -0.173818*Scal.Fl + 0.337115*Scal.Fd1 + 0.348389*Scal.Fd2 where 0 < Scal.fl, Scal.fd1, Scal.fd2 <1	50879.5	96159.2	96174.01
Ficus callosa	ln(Total seeds) = 6.226959 -0.420547*Scal.Fl + 0.132417*Scal.Fd1 + 0.466847*Scal.Fd2 where 0< Scal.Fl, Scal.Fd1, Scal.Fd2<1	3524.97	19252.9	19267.81

TABLE II POISSON MODELS FOR AREAS

Area	**Model**	**MSE**	**AIC**	**BIC**
Kandy Municipal	ln(Total seeds) = 8.13141 -1.47724*Scal.Fl - 0.44274*Scal.Fd1 - 0.22125* Scal.Fd2 where 0< Scal.Fl, Scal.Fd1, Scal.Fd2<1	83372.05	29181.02	29185.64
Thumpane	$\ln(Total\ seeds) = 5.63144\ + 1.235077 * Scal.Fl\ + 0.112589 * Scal.Fd1 +\ 0.985966 * Scal.Fd2$ where 0 < Scal.Fl ,Scal.Fd1, Scal.Fd2 < 1	69983.65	36547.25	36552.25
University of Peradeniya	$\ln(Total\ seeds) = 5.346 - 2.252 * scal.Fl + 1.654 * Scal.Fd1 + 3.229 * Scal.Fd2$ where 0 < Scal.Fl ,Scal.Fd1, Scal.Fd2 < 1	29055.89	94707.34	94711.01

Multicollinearity was not observed (VIF<10) in that model. Poisson models for predicting the yield of seeds of *Ficus racemosa* is the best model with same scaled variables, but mean squared error value is much larger. The results are represented in TABLE I. Moreover, TABLE II demonstrates the results for Poisson models for areas.

In generalized additive modeling approach when models are fitted to areas by merging two species together it was given better results, with least mean squared error values of testing samples. These differences between observed and predicted values verses observed values are graphically represented in Figure1.

The yield of seeds in Thumpane and Kandy municipal area were best described by models which fruit length as the only predictor with moderately large R^2 values (0.599, 0.652) and lower mean square error values for testingsamples. The best model for University of Peradeniya area was given much lower R^2 (0.371) value and higher mean square error value using fruit length and two diameters as predictors. TABLE III gives generalized additive modeling (GAM) results for areas.

All GAMs described are belonging to the family of Gaussian and identity link function. It adds some complexity to the model. Here "s" (s_1, s_2) defines a smoothing function when equations are tabulated as in TABLE III. In local polynomial regression method, curves have been constructed for fruit length, fruit diameter1, and fruit diameter2 as single predictors with optimal bandwidths (0.8109, 1.2938, 0.6045) as in Figure2, 3 and 4 respectively.

TABLE III. GENERALIZED ADDITIVE MODELS FOR AREAS

Area	**Model**	**AIC**	**R^2 value**
Kandy Municipal	Total seeds = 1447.5 + s_1 [Fruit Length,6.985]	3160.4	0.652
Thumpane	Total seeds = 1447.48 + s_2 [Fruit Length,8.65]	3173.1	0.599

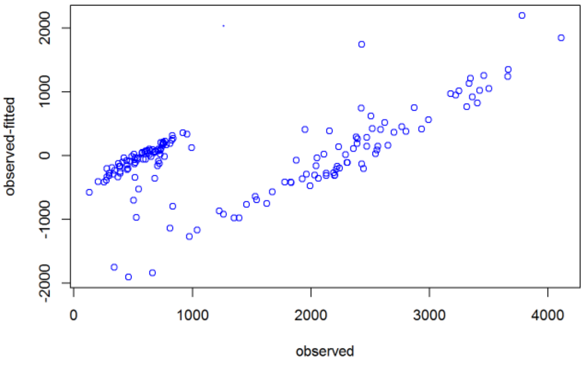

Figure 1. Variation of difference between observed and fitted values with observed values for the generalized additive model in Kandy municipal area

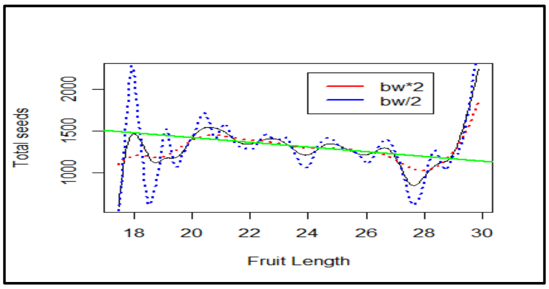

Figure 2. Local Polynomial regression model for a yield of seeds with fruit length.

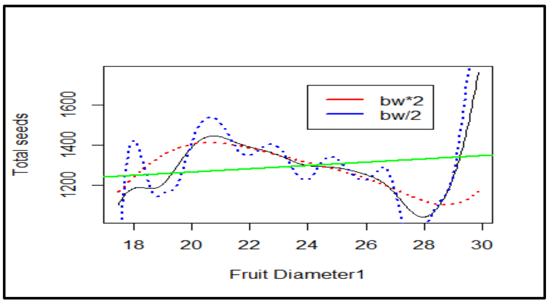

Figure 3. Local Polynomial regression model for a yield of seeds with fruit diameter1.

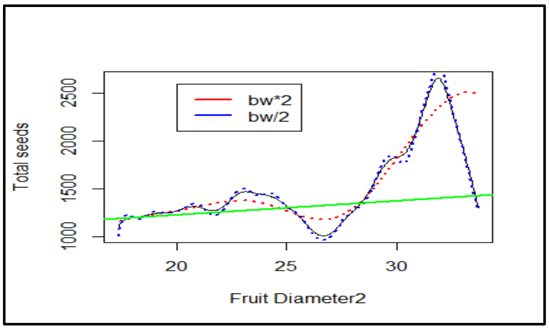

Figure 4. Local Polynomial regression model for a yield of seeds with fruit diameter 2.

It has been shown how it affects increasing (red dotted line, bandwidth*2) or decreasing (blue dotted line, bandwidth/2) bandwidth for biasedness and variance. The descriptive measurements of fruit length show that a large variation of fruit length in Kandy municipal area that also gives best predicting model.

IV. Discussion & Conclusion

In our study, *Ficus* fruits' yield of seeds has been predicted statistically using GAMs and local polynomial regression under the basis of nonparametric quality. Also, Poisson regression technique was applied. In this study, fruit dimensions have been used for predicting the yield of seeds. In order to minimize the effect of external factors, same aged trees and sampling areas with approximately equal climatic features were used. Seeds were separated from damaged and undamaged seeds with the purpose for further analysis. Building models with dimensional predictors were handled properly after checking multicollinearity of predictors which can inflate variances and decrease precision and predictiveness [10]. At the model validation process, testing samples were used to validate models and AIC, BIC values were used to select the best model in practical situations.

Fruits in Kandy municipal area have covered large variation of fruit dimensions (fruit length) and the seed predicting model built for Kandy municipal exhibited more accurate predictions. The combination of two results showed that higher variation of fruit length gives the most accurate result of the yield of seeds. Moreover, the yield of seeds of *Ficus callosa* and *Ficus racemosa* can be predicted using only fruit length and diameter1 which implies the presence of two diameters are not essential for predicting seeds when models are built up using generalized additive models. Poisson model for *ficus callosa* gave more accurate predictions than GAM. When constructing models for sampling areas, fruit length is used as a predictor for Kandy municipal area and Thumpane while two diameters are taken as predictors for University of Peradeniya area. More accurate predictions of GAMs for sampling areas reveal that the type of species doesn't play a major role to predict the yield of seeds for these two species. By plotting difference of observed and fitted values, versus observed values, it implies that when the yield of seeds is less than 1000, fitting values reach to observed values. But there are large deviations when fitting for higher yield of seeds. Using graphical methods with different bandwidth values, it can be concluded that local polynomial regression curves with optimal bandwidth are keeping variance and biasedness in a stable manner. Deviation of optimal bandwidth value (increasing or decreasing) result in curves with unstable biasedness and variances.

REFERENCES

[1] M. J. Shanahan, *Ficus* seed dispersal guilds: ecology, evolution and conservation implications. *Ficus seed dispersal guilds: ecology, evolution and conservation implications*, pp.1-10, 2000.

[2] D. H. Janzen, How to be a fig. Ann. Rev. Ecol. Syst., vol.10, pp**.** 13-51, 1979.

[3] I. Karunarathna, Pollinator and non-pollinator fig wasp relationship in syconia of *ficus exasperate*, vol.2, pp.62-73, 2009.

[4] E. Missanjo, D. Ndalwo, and D. Malinga, Stand Age and Diameter Class Effect on Seed Production of *Pinus kesiya* Royle ex Gordon grown in Malawi, vol.2B, pp.173-177, 2015.

[5] A. Zeileis, and C. Kleiber, Regression models for count data in R, pp.5-7, 2008.

[6] R. Imon, *Nonparametric and Robust Regression*, pp.27-29, 2010.

[7] P. Hall, and J. Racine, Infinite order cross-validated local polynomial regression, pp.510-525, 2015.

[8] S. Wood, Generalized Additive Models: an introduction with R, pp. 119-125, 2010.

[9] T. Hastie, and R. Tibshirani, Generalized additive models. *Statistical Science*, vol.3, pp.297-309, 1986.

[10] W. Cai, Fitting Generalized Additive Models with the GAM Procedure in SAS 9.2, pp.1-2, 2008.