# MTH 643A
# COURSE: SPATIAL STATISTICS

# REPORT ON SPATIAL PREDICTION OF WEED INTENSITIES FROM EXACT COUNT DATA AND IMAGE BASED ESTIMATES

**DEPARTMENT OF MATHEMATICS AND STATISTICS**

INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

**UNDER GUIDANCE OF**

Dr. Arnab Hazra

## SUBMITTED BY

Subhadra Kumari (221437) and Sudesh Kumari (221440)

## Acknowledgments

We would like to express our sincere gratitude and appreciation to everyone who has contributed to the successful completion of this report on spatial prediction of weed intensities from exact count data and image-based estimates.

We would also like to express our gratitude to Prof. Arnab Hazra for his valuable guidance and support throughout the duration of this project. His expertise and insights have greatly contributed to the quality and depth of our analysis.

Furthermore, we would like to acknowledge the resources and materials provided by IIT Kanpur’s spatial statistics course (MTH643A). The knowledge and skills gained from this course have been instrumental in our understanding of the subject matter and the development of this report.

Lastly, we would like to thank our families and friends for their encouragement and support during the course of this project. Their unwavering support has been invaluable.
Once again, we express our heartfelt appreciation to everyone involved in this endeavor.

Thank you for your contribution and support.

Sincerely,
Group-8

## Abstract

The aim of this report is to obtain a weed count in the agriculture field using image analysis. Weed count collection in agricultural fields is traditionally a time-consuming task. However, leveraging image analysis algorithms for object extraction from field images offers a high-resolution (approximately 1 m2) estimation of weed content.

By acquiring images from various field locations, cost-effective weed content maps for the entire field can be generated. Nevertheless, these image-based estimates are not flawless, and obtaining precise weed counts remains invaluable for both validating the accuracy of image-based algorithms and enhancing the quality of estimates through data integration.

In this study, we propose and compare different models for image indexing and exact weed count collection. Our aim is to determine how to effectively combine these data sources to produce reliable field maps. To demonstrate the practicality of our approach, we apply it to real data obtained from a 30-hectare agricultural field.

Our findings indicate that incorporating image-based estimates alongside exact counts significantly enhances map accuracy. Furthermore, we observe that the relative performance of the methods is contingent on the dataset size and the methodology employed (full Bayesian versus plug-in).

**Keywords**: Approximate Cox process; Gaussian random field; Image analysis; Model-based geostatistics; Multivariate data; Poisson regression; Precision farming; Spatial prediction

# Contents

# 1 Statistical approaches for weed mapping

Due to growing environmental concern and not to spread the necessary chemical product in fields , this concern has prompted farmers to check their field's soil and vegetation characteristics on a small scale.In the middle of the fields, where crops and annoying weeds were fighting, farmers had a big problem. They wanted to save their crops from those stubborn weeds, but the usual way was to use chemicals called pesticides. The trouble was, these chemicals could harm the environment and the land they were trying to take care of.

Farmers really care about the environment, so they're coming up with smart ideas to keep it safe. They noticed that using too many chemicals, especially herbicides, is not good for the land. So, they've started doing things differently.
To use herbicides effectively it's important to know where weeds are located in the field. That means we needed to have a detailed map showing where the weeds are located, and this should be very accurate. This kind of map will help farmers to apply the right amount of herbicide in the right places, which is good for both the environment and their crops.
By doing this instead of using the same amount of weed-killing spray everywhere, there can be a idea to use spatial maps to locate these locations in fields. These maps show exactly where the weeds are in their fields, like a super close-up picture. With this information, farmers can decide

how much spray to use in each area. It's like giving medicine only where it's needed.

The statistical approaches that have been developed to obtain such maps may be based on exact weed counts only. The earlier approaches to obtain maps from weed counts have only been based on traditional geostatistical methods.

Implementation of such methods can be found in the works of Heisel et al. (1996), Rew and Cousens (2001), Rew et al. (2001) and Dille et al.(2002), where it is reported that simple interpolation procedures often behave poorly for weed mapping when the number of samples that are available is restricted by economical constraints (fewer than 10 data points per hectare).

The other method shows that the Use of the exact weed count is dependent on the other variables that are correlated with weed like it will depend on the organic matter in the soil, remote sensing estimates have been considered.

But the improvement compared with the use of exact counts depends only on the strength of the relationship between exact counts and covariates, which is usually quite low.

Strategies that are based on analysing pictures of the ground at small scale (about 1 m2) to obtain estimates of weed (or crop) content have been widely studied; see for example Marchant and Brivot (1995), Brivot andMarchant (1996, 1998), Andreasen et al.(1997), Perez et al.(2000), Soille (2000), Tillett et al. (2001), Åstrand and Baerveldt (2002), Onyango and Marchant (2003) and Gerhards and Christensen (2003).the accuracy of the image estimates at a particular site is assessed by comparing its output with the weed content at this particular site. However, in practice collecting and processing a very large number of pictures can be prohibitive in terms of time, and visiting the whole field at a late stage of growth can damage the crop.Hence it can be informative to collect also exact weed count data to assess the accuracy of the image analysis algorithm, and possibly to combine exact count and image-derived estimates in the interpolation scheme to increase its accuracy.
The aim of the present paper is to propose and implement a statistical model to address these two tasks. we proposed two models for predicting the image estimates and then compared out of them which model is giving best image estimates.

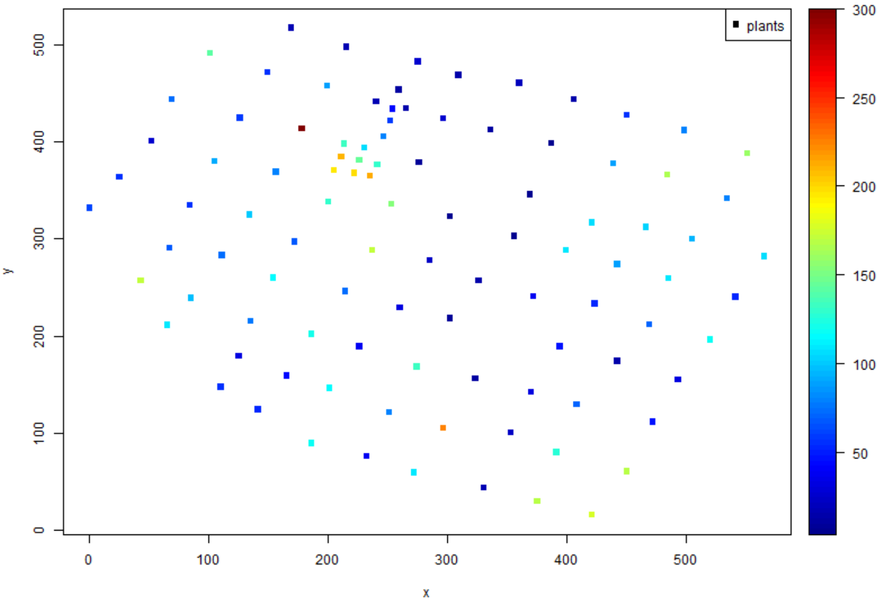


Fig. 1. Spatial distribution of the exact weed count; the colour code denotes the number of observed plants

Figure 1: Exact weed counts.

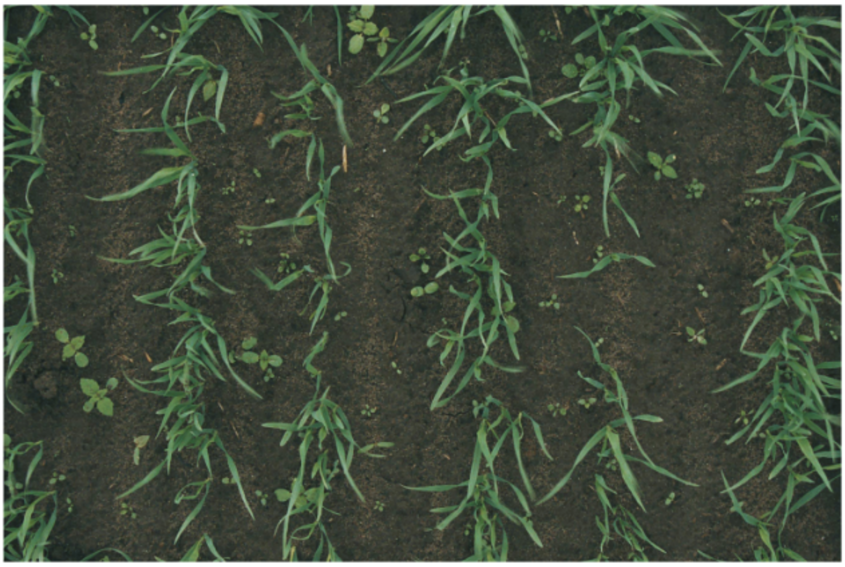
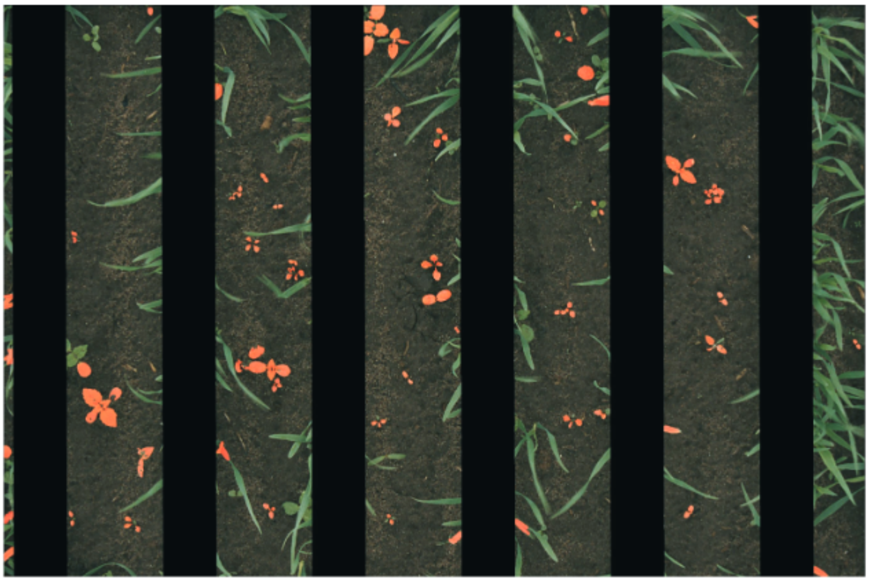
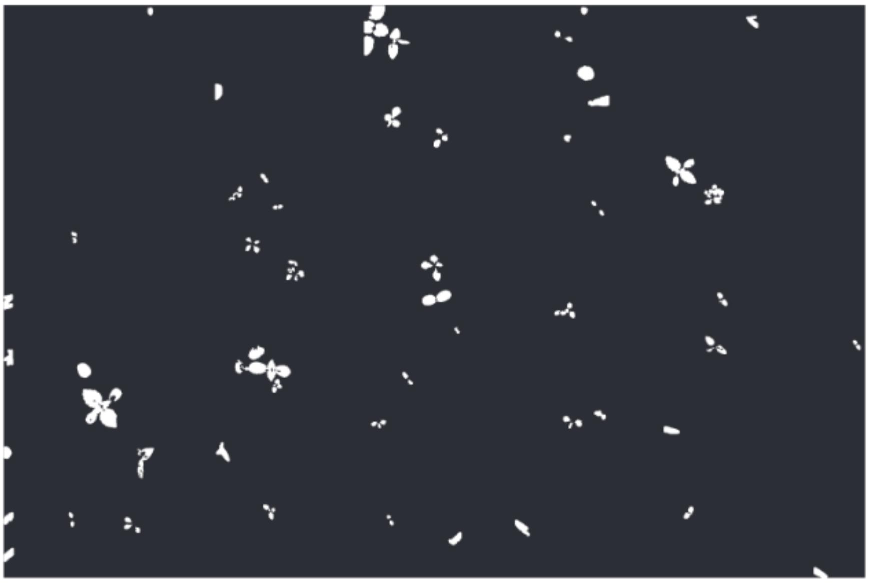


Figure 2: Field picture taken from the original paper.



(b)

Figure 3: Taken from the original paper.



(c)

Figure 4: Taken from original picture.

## 2 A weed data set

Our analysis is based on data that were collected at the Bjertorp farm located 58.26○ N–13.13○ E in the south-west of Sweden. The dataset contains the easting , northing meter and exact weed count and image count as of it's features. The data have been collected in a cultivated wheat field of 30 ha and consist of exact counts in frames of 0.5 m × 0.75 m taken at 100 sites and estimates derived from pictures that were taken exactly over these frames. in this various species were observed are common chickweed (Stellaria Media), field pennycress (Thlaspi Arvense), field pansy (Viola Arvensis), knotgrass (Polygonum Aviculare) and black bindweed (Fallopia Convolvulus), but we will not make the use of individual species counts and their counts compared with total counts of weed in the field.

For the image weed count the image dataset is taken by the digital camera. The images were recorded at a resolution of 008 pixels × 2000 pixels.

For the image based dataset the algorithm used to detect and count the weeds involves the following steps :
- Soil and the plants were differentiated based on their green transformating at a subsequent thresholds.
- Small objects considered as noise and are removed from the binary image.
- Differentiating between crop and rows detected by using the Hough transform (Marchant, 1996).
- Weed that are located between the crop rows and uncovered by crop were found by using masking (Russ, 1994).
-Large weeds covered by crop straws were extracted by using combinations of morphological operations (Soille, 2000).

**Ranges for the exact weed count and image count**
- The exact counts are integers ranging from 4 to 300.
- Image estimates are discrete and range between 7 and 222.

The spatially structured nature of the weed counts appears more clearly on the empirical variogram of weed counts for various binnings. It displays a decrease beyond 150 m, a phenomenon which is known as the hole effect in geo- statistics hat occurs when high values tend to be systematically surrounded by low values, but a good fit of this empirical variogram at short-to-medium distances is obtained with a theoretical exponential model with a spatial scale parameter equal to 50 m.

We plotted the histograms for the exact counts and the image estimates and also pair plots of these two variables. Although these histograms of spatially correlated data can be misleading, they suggest that a crude Gaussian assumption would not be suitable.

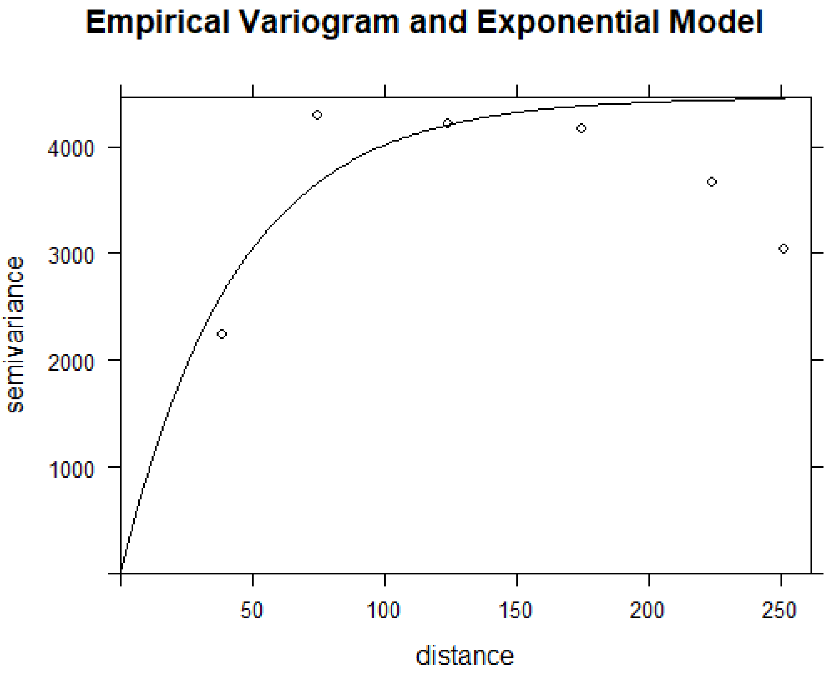


Figure 5: Empirical variogram of the exact count data c(s), and exponential model.

# 3 Modelling

Exact weed counts are observed at a set of sites S = $(s_1, s_2, s_3, ...., s_{n_s})$.And denoted by $C_S$ = $(c(s_1), c(s_2), c(s_3), ...., c(s_{n_s})$. Image estimates are given at a set of sites T = ( $t_1$, $t_2$, $t_3$, $t_4$, ..., $t_{n_t}$). And denoted by $i_T = (i(t_1), i(t_2), i(t_3, ...i(t_{n_t})$. We also denote by $U = (u_1, u_2, u_3), ...., u_{n_u}$ the set of sites at which spatial prediction of the weed field c is sought. for our dataset S and T are considered identical.

## 3.1 Model for exact count

We are considering two models where the spatial variations are essentially captured by a Gaussian random field.

### 3.1.1 Transformed Gaussian random-field model

A first simple model accounting for spatial auto-correlation and non-Gaussian marginal distribution is the transformed Gaussian random-field (TGRF) model.
In this model it is assumed that there is a zero-mean, unit variance, Gaussian random field y(s), $s \in \mathbb{R}^2$. And also considered a function $\psi : s \in \mathbb{R} \to \mathbb{R}$ such that $c(s) = \psi(y(s))$. Such consideration makes it possible to account for non-Gaussian marginal distributions while keeping the parsimony and flexibility of Gaussian random fields for the modelling of spatial variations.
Also We assume that y has a stationary and isotropic correlation function $\rho(h)$ , with exponential decay, namely

$$corr(y(s), y(s+h)) = exp(-||h||/\kappa) \quad (1)$$

where $\kappa$ is unknown correlation parameter that we need to estimate it.
Although the spatial correlation structure is formulated in terms of y whereas the empirical variogram that is available is that of c it is known (see for example De Oliveira (2003)) that, for a large class of continuous transformations $\psi$, the variograms of the latent Gaussian field and of the observed field are quite similar.
Hence, it is expected that the exponential model will give a good fit, at least for short to medium distances.

### 3.1.2 Approximate transformed Gaussian Cox process

For the TGCP we are considering it as the continuous distribution of c but actually here it is a discrete variable. For this model it is proposed to consider that the previous transformed Gaussian field defines an intensity field and that, conditionally on $w = \psi(y)$.
ie.

$$c(s)|w(s) \sim poisson(w(s)) \quad (2)$$

This turns out to be an approximate Cox process with transformed Gaussian intensity and a special case of a spatial generalized linear model.
We named this model as a transformed Gaussian Cox process (TGCP).
we note that under this model with Poisson regression relating c and w, we have

$$cov(c(s), c(s+h)) = E[cov(c(s), c(s+h)|w(s), w(s+h)] + cov(E[c(s)|w(s)], E[c(s+h)|w(s+h)]) \quad (3)$$

and it given w(.) and c(.) are independent so it becomes.

$$cov(c(s), c(s+h)) = 0 + cov(w(s), w(s+h)) \quad (4)$$

and ,

$$var(c(s)) = E[var(c(s)|w(s)] + var(E[(c(s)|w(s))] = E[w(s)] + var(w(s)) \quad (5)$$

The $\psi$, transformation that is considered , we considered this functions such that $\phi(y)/$ has gamma or log-normal marginal distribution. We denote by $\mu$ and $\sigma^2$ the stationary mean and variance of $\psi(y)$.

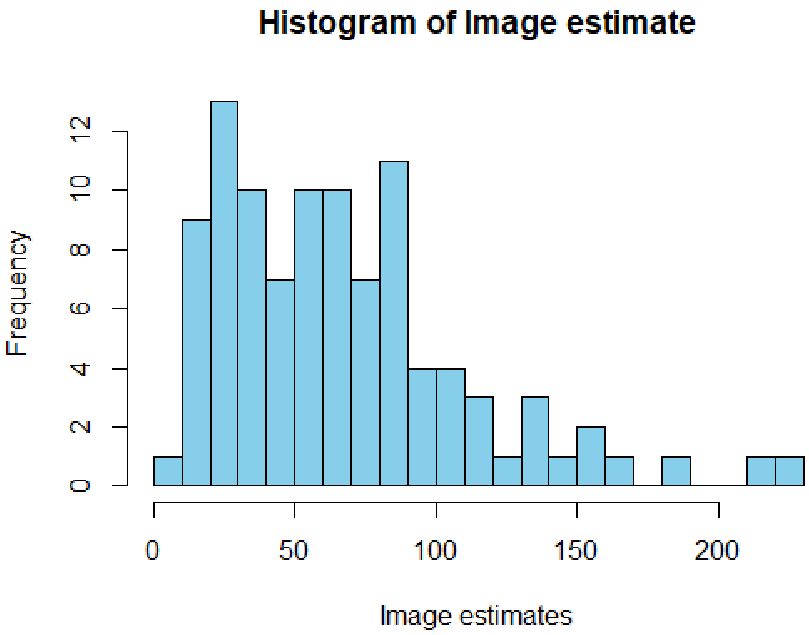


Figure 6: Histogram of Image Estimates.

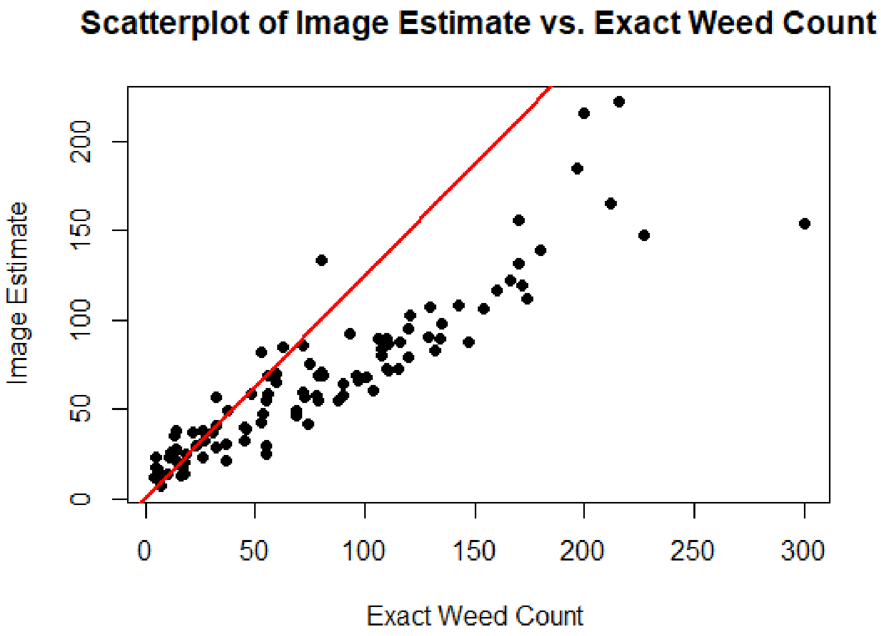


Figure 7: Pair plot of image estimates and exact counts.

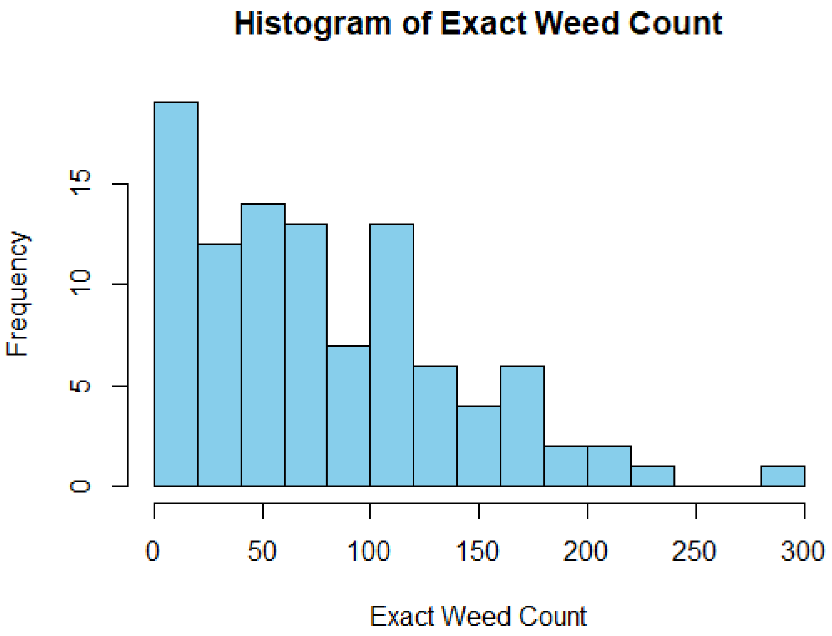


Figure 8: Histogram of exact weed counts.

### 3.2 Modelling the relationship between exact counts and image estimates

For capturing the relation between two variables that is image based weed and exact weed count . We are assuming this is a poisson regression model for the image estimate:

$$i(s)|c \sim poisson(\gamma c(s)^{\delta}) \tag{6}$$

here $\gamma$ and $\delta$ are prameters that are because showing the non linear relation between image estimate and exact weed count. we need to estimate them .

### 3.3 Likelihood formation

so our full inference will be modeled as :
Data that is provided to us is as $i_s = (i(s_1), i(s_2), i(s_3, ...i(s_{n_s}))$. The relations that are provided are :

$$i(s)|c \sim poisson(\gamma c(s)^{\delta})$$

$$c(s)|w(s) \sim poisson(w(s))$$

$$w = \psi(y)$$

$$Y(.) \sim GP(\mu(.), k(s,s') = exp(-||h|/\kappa))$$

A total of parameters that we need to estimate are $\theta = (\mu, \sigma^2, \kappa, \gamma, \delta)$
The likelihood function is given by :

$$L(\mu,\sigma,\kappa,\gamma,\delta) = \pi((i(s_j), c(s_j)), j = 1,2,3,..n; \mu,\sigma,\gamma,\kappa,\delta)$$

$$= \pi(i(s_1), i(s_2) \dots i(s_{n_s})|c(s_1), c(s_2), c(s_3) \dots c(s_{n_s}); \gamma,\delta) \times \pi(c(s_1), c(s_2), c(s_3) .... c(s_{n_s}); \mu,\sigma,\kappa)$$

$$= \prod_{j=1}^{n_s} \pi((i(s_j)|c(s_j)); \gamma,\delta) \times \pi(c(s_1), c(s_2) \dots (s_{n_s}); \mu,\sigma,\kappa)$$

$$\prod_{j=1}^{n_s} \pi((i(s_j)|c(s_j)); \gamma,\delta) \times \int \int \cdots \int \pi(c(s_1) \dots c(s_{n_s})|(Y(s_1), Y(s_2)), \dots Y(s_{n_s})) \times \pi((Y(s_1)) \dots Y(s_{n_s}); \mu,\sigma,\kappa) dY(s_1) \dots$$

$$\prod_{j=1}^{n_s} \pi((i(s_j)|c(s_j)); \gamma,\delta) \times \int \int \cdots \int \prod_{j=1}^{n_s} \pi(c(s_j)|Y(s_j)) \times \frac{1}{|2\pi\sigma^2 \sum_{\kappa})|} \exp^{\frac{-1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{\mu})^T \sum_{\kappa}^{-1} (\boldsymbol{Y}-\boldsymbol{\mu})} d\boldsymbol{Y}$$

This integral is over Y = $(Y(s_1), Y(s_2) \dots Y(s_{n_s}))$

In this likelihood function

$$\pi((i(s_j)|c(s_j)); \gamma,\delta) = \frac{e^{-\gamma c(s_j)^{\delta}} (\gamma c(s_j)^{\delta})^{i(s_j)}}{i(s_j)!}$$

$$\pi((c(s_j)|Y(s_j))) = \frac{e^{-\psi(Y(s_j)} (\psi(Y(s_j)))^{c(s_j)}}{c(s_j)!}$$

## 4 Implementation on real data

#### 4.0.1 4.1.1.Gamma versus log-normal transform

We checked that the gamma and the log-normal distribution gave similar results in terms of goodness of fit for the TGCP and the TGRF model. Hence, we assume hereafter that $\psi(y)$ has a gamma marginal distribution both for the TGRF and the TGCP model.

### 4.0.2 Transformed Gaussian random-field model versus transformed Gaussian Cox process model

We checked the fit of the TGRF model and the TGCP model to data (within the global model including image estimate data) as follows: we plotted the pair plot and the Quantile - Quantile plot of the estimated image versus the exact weed count.

## 4.1 Assessment of prediction accuracy

We accessed the accuracy of the two models by their respective plots of estimated image estimate versus the exact weed count at the same locations.

### 4.1.1 Setting for the transformed Gaussian Cox process and transformed Gaussian random-field models

Pair plot of the exact weed count versus image estimates of using TGRF model.We compared the TGCP model and the TGRF model in terms of accuracy of prediction of weed counts on various subsamples of the Bjertorp data set. We considered 100 random subsamples with various numbers of exact count. we predicted the image counts. Then plotted the pair plot and the Quantile - Quantile plot of Exact weed count versus the image estimates to get an estimate.

### 4.1.2 Accuracy Plots for the TGRF model are :

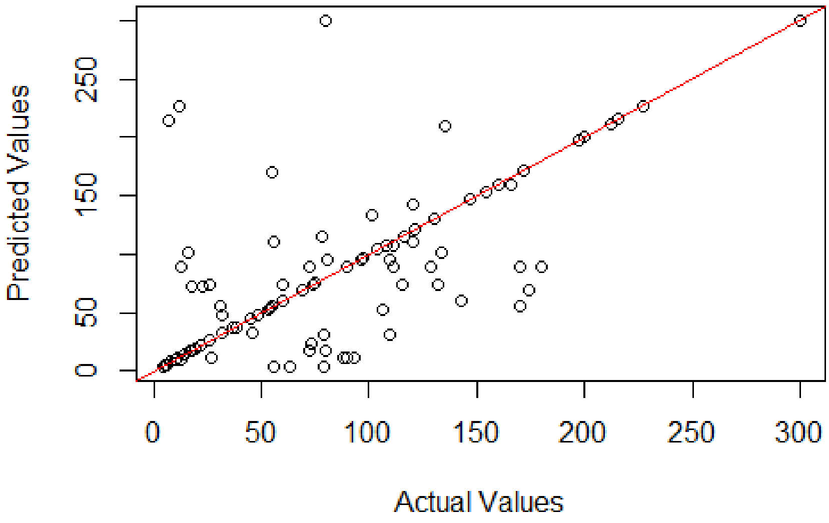


Figure 9: Pair plot of predictive distribution Image Estimates against exact weed count.

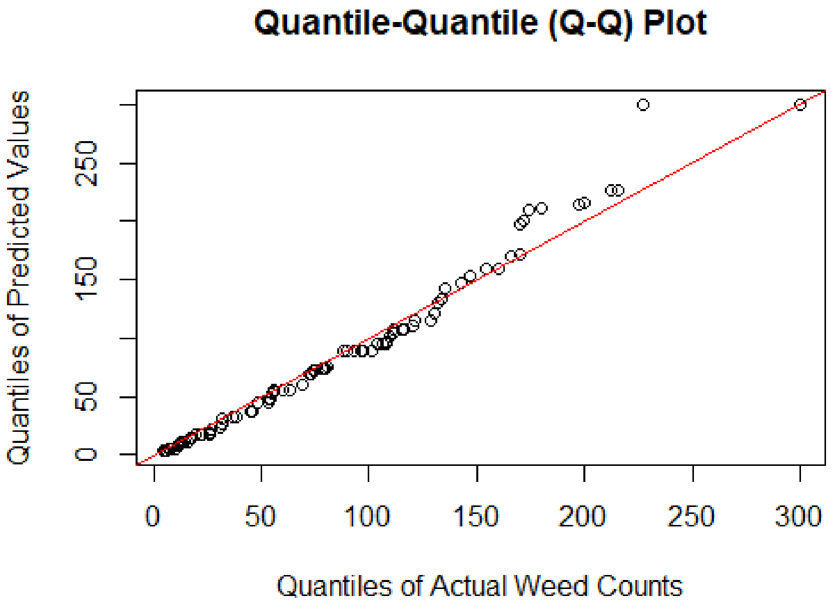


Figure 10: Q-Q plot of predictive distribution Image Estimates against exact weed count for TGRF.

### 4.1.3 Accuracy Plots for the TGCP model are :

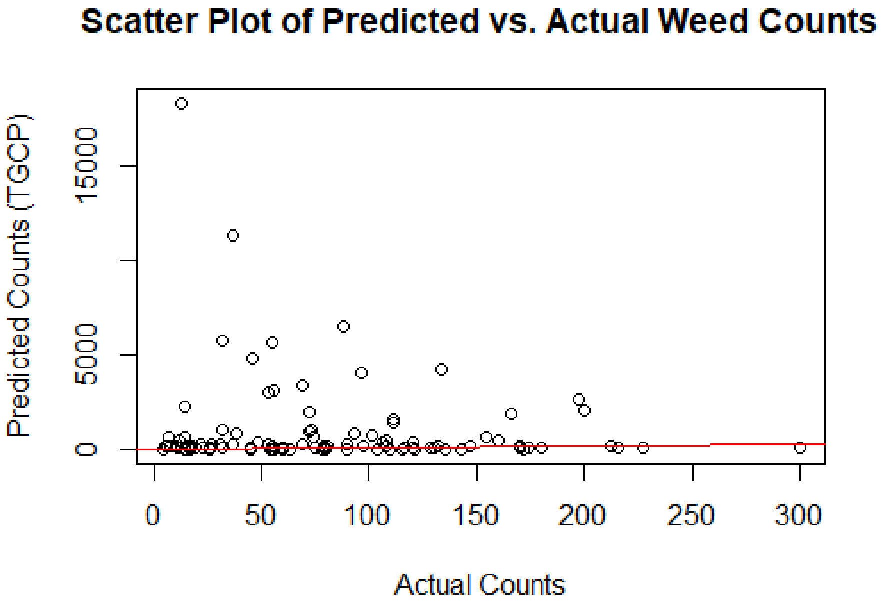


Figure 11: Pair plot of predictive distribution Image Estimates against exact weed count for TGCP model.

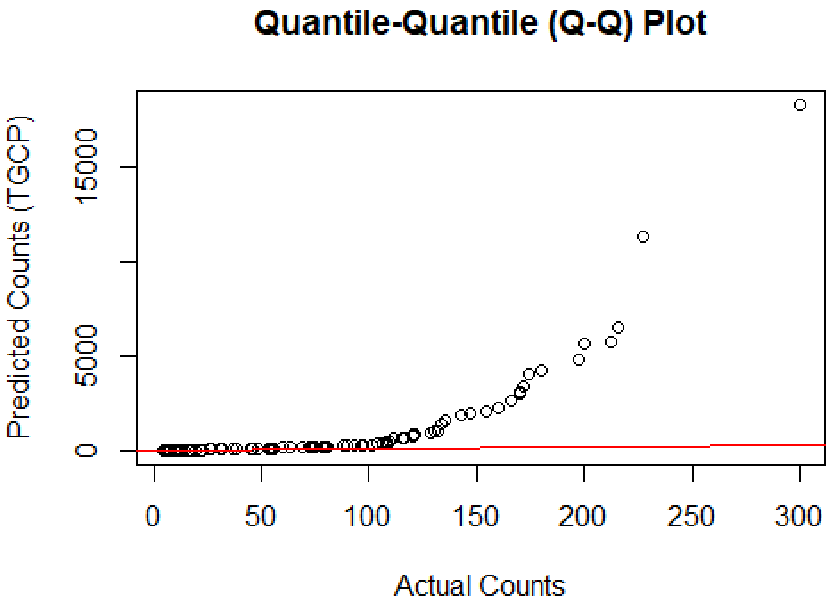


Figure 12: Quantile - Quantile plot of predictive distribution Image Estimates against exact weed count for TGCP model.

### 4.1.4 Methods for estimating the parameters

- For the estimation of the kappa : we used the profile likelihood estimation for the estimation of the the parameter involves in the correlation function (kappa)
- For the estimation of the gamma and delta : For estimating the gamma and delta parameters that are involved in the relation between the image estimates and the exact count, we used probit GLM method.
- For the estimation of $\mu$ and $\sigma$ : For the estimation of $\mu$ and $\sigma$ we used the simple Monte Carlo estimation approach .

### 4.1.5 The estimated values for the parameters

- The value of the $\kappa$, we get from the estimation is 1.13
- The estimated value for the parameter $\mu$ is 2.32
- The estimated value for the $\sigma^2$ is 26.56
- The estimated value for the parameter $\gamma$ is 3.90
- The estimated value for the parameter $\delta$ is 0.725

## 5 Conclusion

1) when we use Image data its clearly improves the accuracy of prediction compared with exact counts. Throughout this study some of the image data were collected at the same sites as exact

counts data. All computations are based on such data.

where $S \cap T = \phi$ that prediction was less accurate than where $S \subset T$ in the TGRF and TGCP model. we also note that there was a poisson regression relationship between image and counts data, where

$$var(i) = \gamma E(c^{\delta}) + \gamma^2 var(c^{\delta}) \tag{7}$$

if $\gamma$ and $\delta$ are close to one
then $var(i) = E(c) + V(c)$ that implies $var(i) > var(c)$
but in the present data $var(i) < var(c)$

2) when we use complex model TGRF and TGCP model on a given small data set TGRF model give more accurate prediction than TGCP model. 3) When we increase the data points then the models will give the results with more accuracy

## 6 References

To conduct a thorough analysis, we've gathered and carefully chosen a variety of reliable sources including research papers and expert articles and the course MTH643 Spatial Statistics resources. These sources will be the building blocks of our research, offering valuable insights and a strong basis for our work.
The main paper that we tried to reproduce can be found at the site:
https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1467-9876.2009.00664.xpane-pcw-figures

## 7 Contribution of teammates

"In our project, we both bring equal dedication and effort, creating success as a team. Our shared contributions make every achievement a joint accomplishment, reflecting our equal partnership in this endeavor."

| Sudesh kumari (221440) | Subhadra kumari (221437) |
|---|---|
| Paper selection | Paper selection |
| variogram plot | exact weed count plot and variogram plot |
| parameter estimation using coding | histogram plots of exact and image data |
| Accuracy plots of the TGCP and TGRF and conclusion | conclusion portion added |
| Prsentation slides (50 percent) | presentation slides (50) percent) |
| Report writing in latex (50 percent) | report writing (50 percent) |
| presentation (50 percent) | presentation (50 percent) |