

9th National Convention on Statistics (NCS)

EDSA Shangri-La Hotel

October 4-5, 2004

**Optimal Selection of Locations at Which to Measure
Factors/Covariates for a Spatial Point Process**

by

Jeffry J. Tejada

For additional information, please contact:

Author's name:	Jeffry J. Tejada
Designation:	Assistant Professor
Agency:	UP School of Statistics, University of the Philippines
Address:	Diliman, Quezon City
Telefax:	(02) 928-0881
E-mail:	jeffry.tejada@up.edu.ph

Optimal Selection of Locations at Which to Measure Factors/Covariates for a Spatial Point Process

by
Jeffrey J. Tejada ¹

ABSTRACT

In the analysis of spatial point processes, one of the usual objectives is the estimation of the underlying intensity measure. Whenever available, information on factors and/or covariates hypothesized to influence the intensity, is incorporated in the analysis. This paper considers the case where, given a realization of a spatial point process, corresponding factors/covariates are yet to be measured. As such, locations in the given region of interest, at which to measure these factors/covariates, may be selected in a manner that is optimal in some sense. The research problem and a proposed solution are discussed in the context of Poisson counts with identity and exponential link functions.

Keywords: spatial point processes, optimal designs, reverse experiments

I. Introduction

A usual aim in the study of spatial data is the characterization of the occurrence of an event of interest over a given region. An example is in disease mapping, where one might investigate the pattern of occurrences of an illness within a certain locality. Another example is the study of the appearances of a particular biological species in a given geographical area, say, a forest or a reef section. Often, the spatial variation of such events may be explained to some extent possibly by factors or covariates. It might be expected, for example, that there are more cases of respiratory condition in areas where pollution level is high, or that locations with larger amounts of a certain chemical in the soil have fewer counts of a plant species growing on them.

Spatial data normally consists of a point pattern realization on a region, or summarized as a collection of counts from disjoint sub-regions. In either case, estimation of the intensity measure or equivalently, the intensity function, usually forms the bigger part of a typical analysis, and is done using any of a number of available methods.

If information on factors and/or covariates that possibly associate with the point process is available, then this may be incorporated in the characterization of the point process. The description and eventual prediction, if desired, of the occurrence of points over the region, become reliable if a large proportion of the spatial variation of the intensity measure can be accounted for by the hypothesized factors and/or covariates.

In this paper, we investigate the situation wherein, given a point process, factor (and/or covariate) information is yet to be obtained. If potential

¹ Assistant Professor, UP School of Statistics, University of the Philippines, Diliman

measurements of these factors are at hand, then some benefits may be realized by purposively selecting where to make these measurements. This “reverse design” may be done in a way that is optimal, say, in that the factor effects are estimated most precisely. This may prove valuable if cost of measurement of particular concern and only a number of these measurements are possible.

An example, albeit extreme, is when the intensity measure is a linear function of one factor, and only two factor measurements are planned. To maximize the information about the parameters of the model, the measurements are made at two different locations, one where the estimated intensity is lowest in the region, and the other, where the estimated intensity is highest. In the respiratory health example, this corresponds to taking pollution levels, say the amount of a certain pollutant, at locations where the concentration of occurrence of respiratory cases is lowest and highest, respectively.

The example uses the criterion of D-optimality, which is the focus of this paper. The goal is to minimize the variance of the parameter estimates, or equivalently, to maximize the precision. This exposition is limited to the case of spatial Poisson point process, whose intensity measure is a function of only one factor. We consider the situation where the data consist of counts, either from an a priori (such as administrative units) or an arbitrary partitioning of the region of interest. The selection of the optimal locations is derived for the linear and log-linear effect models.

II. Reverse Experimental Designs

Let R be a bounded geographical region, possibly with a given partition, e.g., administrative sub-regions. For any subset A of R , let $N(A)$ denote the number of occurrences (termed points) in A , of an event of interest. We assume for every $A \subset R$, that $N(A)$ is Poisson distributed with mean $\Lambda(A) = \int_A \lambda(r)dr$, and for any two disjoint sub-regions A and B , $N(A)$ and $N(B)$ are independent. As it varies over R , the process N is referred to as a spatial Poisson point process with intensity function λ . Since we consider only the case of Poisson counts in this paper, the derivations involve mainly the intensity measure Λ , and not the intensity function λ . We thus use the term “intensity” to mean to the intensity measure.

Suppose x is a fixed process on R such that the intensity Λ evaluated at a sub-region A is a function of $X(A) = \int_A x(r)dr$. We write $\Lambda(A) = f(X(A))$, or simply, $\Lambda = f(X)$ for some function f . Note that $X(A)$ is a characteristic of the entire sub-region A , and not just of a particular location within A .

We consider two such functions by which Λ is modeled, namely:

(a) the first-order linear model,

$$\Lambda = \beta_0 + \beta_1 X, \tag{1}$$

and

(b) the first-order log-linear model, $\ln(\Lambda) = \beta_0 + \beta_1 X$ or

$$\Lambda = \exp(\beta_0 + \beta_1 X). \quad (2)$$

In a forward experimental design problem, factor levels are selected at which to measure the response variable. The criterion of D-optimality requires this selection to result in parameter estimates which have the smallest variance over all possible selections.

In this research, the response is the intensity and the factor is X . Since an estimate, $\hat{\Lambda}$, can be obtained from the realization of N , and information on X has yet to be gathered, we thus have the reverse problem of selecting $\hat{\Lambda}$ -values (and ultimately the locations) at which to measure the X -levels that give rise to these intensity estimates. Using the D-criterion, we obtain reverse designs such that the selection of $\hat{\Lambda}$ -values corresponds to estimates of model parameters which have minimum variance.

Formally, let $\{A_t, t \in T\}$ be a partition of R . Writing $\Lambda_t = \Lambda(A_t)$ and $X_t = X(A_t)$, suppose the realization of the spatial Poisson point process N yields the observation set $\{\hat{\Lambda}_t, t \in T\}$ of estimated intensities. For a given sample size n , we wish to select a subset $S \subset T$, of size n , and measure the factor levels $X_t, t \in S$, which give rise to the subset $\{\hat{\Lambda}_t, t \in S\}$ through the model, $\hat{\Lambda}_t = \hat{f}(X_t)$, where \hat{f} denotes the estimate of f .

The selection is made such that the determinant of the Fisher information matrix, $I(\beta)$, given by the formula,

$$I(\beta) = -E \left(\frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta'} \right) \quad (3)$$

is maximum over all possible selections, where $L(\beta)$ is the likelihood function of the parameter vector $\beta = [\beta_0, \beta_1]'$.

Whereas in a forward design problem, the maximum of the above criterion is found in terms of the factor variable, for the reverse problem posed here, the proposed solution is to find the maximum of the criterion for values of the response, namely, the intensity measure.

II.1 D-Optimal Reverse Design for Identity Link Model

We now derive the D-optimal reverse design for the Poisson regression model $N \sim \text{Po}(\Lambda)$ with intensity $\Lambda = \beta_0 + \beta_1 X$. To estimate the coefficients, let Λ_1 and Λ_2 be the support points of the design, and let p_1 and p_2 be the proportion of the sample size allocated to Λ_1 and Λ_2 , respectively.

The Poisson log-likelihood function for a sample of size n is given by

$$\ln L(\beta) = \sum_{i=1}^2 np_i N_i \ln \Lambda_i - \sum_{i=1}^2 np_i \Lambda_i - \sum_{i=1}^2 np_i \ln N_i! \quad (4)$$

where $\Lambda_i = \beta_0 + \beta_1 X_i$, $i = 1, 2$.

The Fisher information matrix is derived as

$$I(\beta) = \begin{bmatrix} \sum_{i=1}^2 \frac{np_i}{\Lambda_i} & \sum_{i=1}^2 \frac{np_i X_i}{\Lambda_i} \\ \sum_{i=1}^2 \frac{np_i X_i}{\Lambda_i} & \sum_{i=1}^2 \frac{np_i X_i^2}{\Lambda_i} \end{bmatrix}. \quad (5)$$

We obtain the determinant of (4) to be

$$\det(I(\beta)) = \frac{n^2 p_1 p_2 (\Lambda_1 - \Lambda_2)^2}{\beta_1^2 \Lambda_1 \Lambda_2}. \quad (6)$$

Letting $h = \frac{\Lambda_1}{\Lambda_2}$, the D-criterion reduces to

$$\det(I(\beta)) = \frac{n^2 p_1 p_2 (1-h)^2}{h}. \quad (7)$$

The maximum is therefore attained by setting $p_1 = p_2 = 0.5$, and taking h to be as far from zero as possible. This means taking the lowest and highest intensity values as the optimal reverse design points, allocating n measurements equally between the two Λ -values.

II.2 D-Optimal Reverse Design for Logarithmic Link Model

The natural link function for Poisson regression is the logarithmic link giving the model $\Lambda = \exp(\beta_0 + \beta_1 X)$.

For this log-linear model, the D-optimal forward design depends on the values of the parameters. Either a locally D-optimal design, which requires the knowledge of or a guess at the parameter values, or a Bayesian design, which requires a prior distribution of the model coefficients, is employed. However, the D-optimal reverse design does not involve such requirement. We now derive it here.

As in section 2.1, let p_i be the proportion of the sample allocated to the design point Λ_i . The D-optimal reverse design has two support values, which we denote by Λ_1 and Λ_2 .

Given a sample size of n , the likelihood function is derived as

$$\ln L(\beta) = \sum_{i=1}^2 np_i N_i \ln \Lambda_i - \sum_{i=1}^2 np_i \Lambda_i - \sum_{i=1}^2 np_i \ln N_i! \quad (8)$$

where $\Lambda_i = \exp(\beta_0 + \beta_1 X_i)$, $i = 1, 2$.

The Fisher information matrix is given by

$$I(\beta) = \begin{bmatrix} \sum_{i=1}^2 np_i \Lambda_i & \sum_{i=1}^2 np_i \Lambda_i X_i \\ \sum_{i=1}^2 np_i \Lambda_i X_i & \sum_{i=1}^2 np_i \Lambda_i X_i^2 \end{bmatrix}, \quad (9)$$

yielding the D-criterion

$$\det(I(\beta)) = \frac{n^2 p_1 p_2 \Lambda_1 \Lambda_2}{\beta_1^2} \log^2 \left(\frac{\Lambda_1}{\Lambda_2} \right), \quad (10)$$

which is maximized when $p_1 = p_2 = 0.5$ and when Λ_1 is the highest intensity value and $\Lambda_2 = 0.1353 \Lambda_1$.

The D-optimal reverse design, therefore, allocates half of the trials at the highest intensity, which we write as Λ_H , and the other half at $0.1353 \Lambda_H$.

III. Optimal Selection of Locations

The D-optimal designs derived for the reverse problem involve only two design values, the same number as in the forward problem. In the latter, replicates are obtained at these factor levels for sample sizes greater than two.

More specifically, if n ($n > 2$) is even, we have $\frac{n}{2}$ replicates each for the two factor levels in the forward case. However, in the reverse case, if we measure X at the same location, we will obtain the same value due to the fact that the factor X is a fixed process. It is the response Λ that is measured with error, owing to N

being a random process. It is therefore unfortunate that in this reverse problem, no purposive replication is feasible.

Thus, the allocation of, say, k trials to a design value Λ^* , is performed, here, to be approximately the selection of k sub-regions whose estimated intensities are closest to Λ^* .

Using the results of section 2, we suppose that spatial data, in the form of counts, are available, either from a given subdivision of the region R , or from an arbitrary partitioning. Assuming an estimator for Λ (which can be the count, N , itself) can be performed, let $\hat{\Lambda}_L$ and $\hat{\Lambda}_H$ the lowest and highest intensity estimates among $\{\hat{\Lambda}(A_T), t \in T\}$ where $\{A_T, t \in T\}$ is a partition of R .

Suppose the sample size n is even. Under the linear model hypothesis, the selection proceeds by taking $\frac{n}{2}$ sub-regions whose intensity estimates are closest to $\hat{\Lambda}_L$, and the other $\frac{n}{2}$ sub-regions whose intensity estimates are closest to $\hat{\Lambda}_H$. Measurements of X then made at these n selected sub-regions. It is emphasized that a measurement X at A , $X(A)$, is a characteristic of the entire sub-region A , such as a total or a maximum amount, and not just that of a particular location in A .

If n is odd, say, $n = 2k + 1$, we select the first $2k$ sub-regions as in the even sample size case, and select the last one to be that sub-region whose estimated intensity is closest to either $\hat{\Lambda}_L$ or $\hat{\Lambda}_H$.

For the log-linear model, the selection procedure is the same as in the linear case, having as design values $\hat{\Lambda}_H$ and $0.1353 \hat{\Lambda}_H$.

IV. Conclusion and Recommendations

Benefits in terms of precision in estimating spatial point process factor effects can be attained by carefully selecting where to conduct factor measurements. Minimum variance of the parameter estimates is obtained by selecting those locations that correspond to optimal design theory, derived for the reverse experiment situation.

More refined selections may possibly be realized when data is in the form of a spatial point pattern. Extensions to more than one factor situation is desired, leading to a greater percentage of the variation of the point process being explained. Instead of using the intensity measure, an analysis using the intensity function, if feasible, might bring about better designs.

References

- Atkinson, A.C., and A. N. Donev. Optimum Experimental Designs. New York: Oxford University, 1992.
- Diggle, P.J. Statistical Analysis of Spatial Point Patterns. London: Academic Press, 1983.
- Moller, J., and R. Waagepetersen. "Statistical Inferences for Cox Processes." Spatial Cluster Modelling. (eds. D. Denison and A.B. Lawson). Chapman and Hall, 2001.
- Rao, C. Radhakrishna, and Helge Toutenburg. Linear Models: Least Squares and Alternatives. 2nd Ed. New York: Springer-Verlag, 1999.