ON A NEW PROBABILISTIC METHOD FOR EXACT SPATIAL INTERPOLATION OF SAMPLING DATA: APPLICATION TO OLIVE...

Markos Avlonitis
Ionian University

ON A NEW PROBABILISTIC METHOD FOR EXACT SPATIAL INTERPOLATION OF SAMPLING DATA: APPLICATION TO OLIVE FRUIT FLY POPULATION DATA

The paper introduces a new probabilistic algorithm for spatial interpolation of sampling data. The algorithm lies on new finding in literature concerning scaling behaviour of complex systems. It is demonstrated that there is a critical spatial scale over which statistical behaviour of local subsystems coincides with the statistical behaviour of the entire system. This observation in combination with local spatial statistical trends is used in order to construct a more efficient algorithm for interpolation of spatial sampling data. Application of the proposed algorithm to real population data of olive fruit fly is demonstrated. Simulation results and evaluation procedures in order to test the robustness of the proposed algorithm are also presented.

1 Introduction

For accurate population census, the ideal process would be to count all the individuals in the population. Since it is impractical to do so, the processing of sampling species numbers is utilized to monitor variations and evolution of a population in space and/or time.

In cases where population sampling is insufficiently dense or is quite dense only in a limited area, in order to describe the evolution of the species, interpolation or extrapolation is normally used to "predict" the missing information. Interpolation, being a local measurement, takes into account an area around the missing sample and combines the samples in this area, based on some interpolation method (Shepard 1968, Burnett 1987, Piegl 1991), to retrieve the new sample value. Such locality though may sometimes fail to grasp global population trends.

More over, regularity of sample network is an important issue since there are a lot of realizations where regularity is not possible to archive. Interpolation from irregularly spaced data occurs in many fields such as in ecology, meteorology, geology and cartography where data are usually sampled from irregularly located observation stations.

On the other hand, more and more dynamical methods that are based on scaling argument emerge in literature in order to improve description of complex systems; see for example Avlonitis et al (2007) and references therein. In a few words, scaling arguments states that some times it is possible to extract information about the dynamic behavior of the entire (arbitrary) system by monitoring its behavior to a local area and vice versa .

To our know length, while scaling argument was applied in order to improve extrapolation algorithms (Peters and Herrick, 2004), there is no such work addressing the interpolation problem. It is the aim of the present paper to explore such possibility. Indeed, it is proposed an explicit probabilistic algorithm where the information of the distribution of the entire system is used in order to achieve a better interpolation method in local missing data.

Treating scaling behavior of dynamic systems is a very difficult task since one must be able to consider appropriate scales of interactions. Short and long range of interactions may distinguish and be responsible for neighborhood effects and global trends correspondingly. Length scales may be determined theoretically or estimated by appropriate signal features extraction techniques.

In the present study the proposed algorithm taking into account scale effects is applied to insect population dynamics. In the study of insect population evolution, locality is important as insects travel beyond their place of birth in search of a more suitable environment to breed and mate. On the other, global trends are present since generation and reproductions are functions to the climate and environmental conditions remaining constant to higher length scales.

More specifically, the work addresses the problem of correct monitoring the spatial population of olive fruit flies by the use of low number of traps in order to predict in time outbreaks (in Corfu island in Hellas). It has been calculated that olive fruit outbreaks results to economic losses up to 15% of the olive crop (Kapatos and Fletcher, 1983). Pesticide treatments are usually applied every year to control the fly population. The use of expensive chemicals and application machinery while increases production cost as well as creating many environmental problems (Bueno and Jones, 2002), most of the time failed to control the population because of asynchronous treatment.

There is a need for more selective methods for the olive fly population spatial monitoring. Towards this goal we attempt to improve prediction of olive fruit fly population spatial arrangement with a novel probabilistic methods taking into account both local and global statistical trends.

2 The new proposed algorithm

The main idea which is applied in this work based in a study of scale behaviour of the dynamic system under consideration. This idea it is not new. Indeed, recently a similar treatment of scalability was applied in the field of computer virology in order to predict the evolution of worm population in the Internet (Avlonitis et al, 2007). The main contribution of this idea of scaled evolution lies in the assumption that there is a large class of dynamical systems which evolution coincides in different length scales. In Avlonitis et al (2007) it was proved that this leads to the conclusion that there is a critical spatial size over which the statistical evolution coincides with that of the entire space. In other words, given a random variable with known (or estimated) distribution function there is a critical length for which the same distribution function describes correctly its statistical behaviour in space (or time).

The above result can be used for spatial interpolation of sampling data as follows. For a given spatial space under examination take a quite dense network of sampling points. Find the distribution function of the sampling data. Repeat the procedure for increasing sampling network size until the same distribution reveal for the first time. This estimated critical size coincides with the desired critical sampling area.

To proceed further, split the entire space under examination in parts with the desired critical sampling area. For each sampling area it is possible to use a less dense sampling network and predict (with interpolation) the rest of the sampling data noting that the observed values (from sampling points) combined with the predicted values must reproduce the estimated distribution function. This extrapolation procedure is expected to be more powerful from other known procedures since the strong restriction of choosing values from a known distribution eliminates statistical errors in prediction.

In order to take into account locality possibly emerged in the spatial data an extra filter of choosing values from the known distribution is applied. Indeed, the interpolate value must not exceed a predefined step value. The estimation of the step barrier can be found from statistical analysis of the original dense sampling network. Details of the algorithm described above are demonstrated below for the case of sampling of population data of olive fruit fly.

3 Application to olive fruit population data: materials and methods

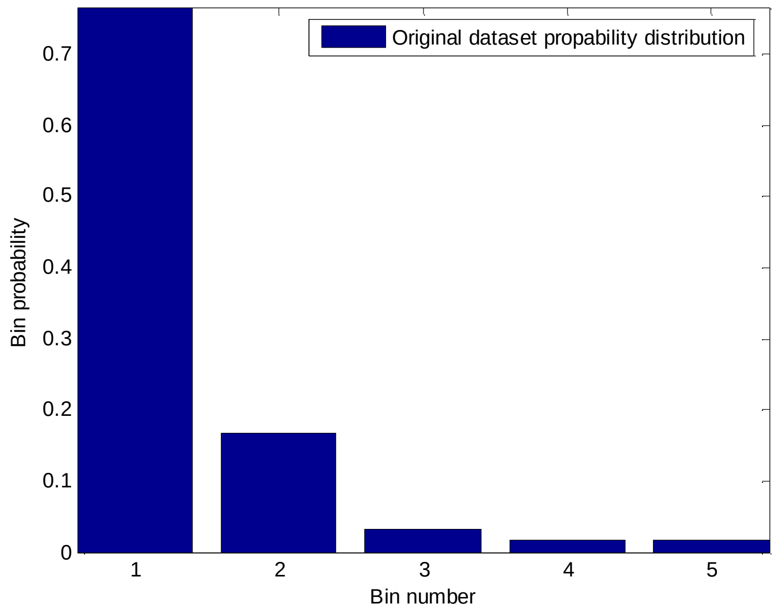
The island of Corfu is located to the north-west of Hellas. The Mediterranean climate type is characterized by high temperatures and very low rainfall during the summer months. The island mostly is covered with olive trees of the variety "lianolia", which possess a marked biannual cycle of fruit production: in many areas on the island a season of high fruit production is followed by a season of low fruit production. The result of this is that a high degree of spatial heterogeneity of olive fruit fly (or D. oleae) is observed since large numbers of adult D. oleae emerge in late May and June in areas where very few fruits will be available for attack during the next season. As a result dispersion of these flies takes place in order to find new oviposition sites (Fletcher and Kapatos, 1981).

In order to monitor the evolution of population in space glass McPhail traps was used. The monitoring of dacus population is of crucial importance since above a critical level chemical methods (or others) are applied in order to prevent population outbreaks which results to the high reduction of production. As it was mentioned above, the traps used were glass McPhail traps which were charged with approximately 100 ml of a 2% solution Entomogyl protein hydrolysate just prior to the releases and then every time the traps were checked (Kapatos et al, 1977). The insects caught in the traps were sieved from the liquid, placed in numbered containers and then taken back to the laboratory where the specimens of D. oleae were separated from the other species, sexed and counted.

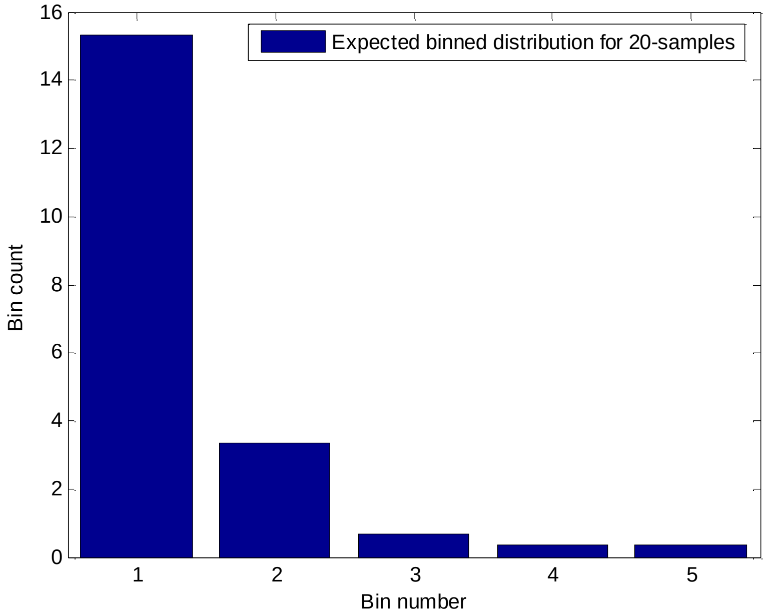
The study area was the municipality of S. George at the north of the Island. The experiment was carried out in June/July, in 2007. The trees were 5 to 8 m high and were planted close together (10 to 12m) forming a homogeneous habitat discontinuing from large areas of living units and lawns. A total number of 60 traps were arranged around the study area the mean distance been of the order of 300m. The arrangement of traps was done in the sense that a quite dense network is formed. The spatial interpolation problem under consideration is as follows: how one can fill up the entire space with accurate values of the population of olive fruit fly when a less dense sampling network is used. The solution to this problem has many practical implications since sampling networks may be simplified and treated with little effort in many ecological areas where spatial signals are monitoring. In the next section the robustness of the proposed method for extrapolating spatial data of olive fruit fly population is demonstrated.

4 Simulation results and evaluation

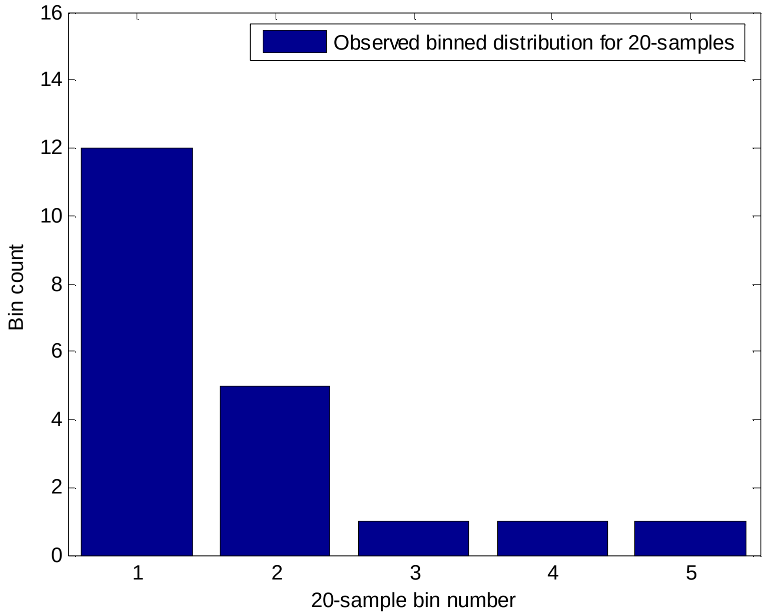
We have acquired a dataset of 60 olive fly samples from 60 olive fly traps placed in an olive grove in Corfu. The sampling field was heterogeneous in the sense the olive tree grove contained bald patches of relatively large size. The olive fly traps where sampled on the 7th of June 2007, a season of peak population growth, just before the annual pesticide treatment commenced.



(a)



(b)



(c)

Figure 1 - Binned probability distributions for the original 60-sample dataset (a), the expected distribution of the 20-sample dataset (b) and observed distribution of the 20-sample dataset (c).

To test the hypothesis that the total population distribution matches the distribution of a smaller, independent subset, we consider a sample set of 20 Olive Fly traps placed along an almost straight line spanning a distance of 6km.

Figure 1a displays the distribution of the original, 60-sample, dataset and 1b the distribution of the 20 sample subset. The samples were grouped in bins of size 5 to make comparisons more meaningful.

We used the chi-squared test to assert that the 20-sample dataset follows the original dataset distribution. Considering the observed and the expected values of the bins we have the following table (Table 1):

Observed bin count	Expected bin count
12	15.333
5	3.333
1	0.666
1	0.333
1	0.333

Table 1 – Observed and expected bin counts.

The calculated chi-squared value is 4.3913. For 4 degrees of freedom (the number of bins minus 1) the chi-squared threshold for 0.05 significance level is 9.448. So it appears, with large confidence, that the 20-sample dataset does indeed follow the original 60-sample dataset distribution.

In order to test the proposed algorithm, we subsampled the 20-sample dataset omitting every other sample. We derived thus a 10-sample dataset from which we will attempt to restore the original 20-sample dataset.

Further, in order to reveal the local gradient behavior of sampled data, we differentiate the samples and for selected sample values we plot the corresponding gradients (Ahnert and Abel, 2007). This produces a sparse point cloud. To estimate the average gradient value, we fit a linear function $g(x)$ to the sparse point cloud. This function will provide an estimate of the gradient for a given sample value x (figure 2).

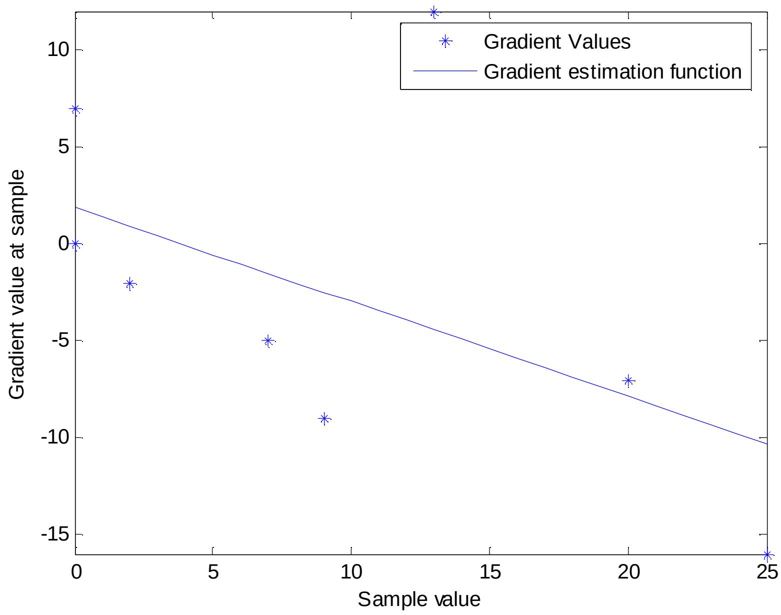


Figure 2 - The gradient of sample values x and the fitted estimation function $g(x)$

Finally we calculated the original 60-sample dataset discrete probability density function. In a good approximation, it appears to follow an exponential-like distribution (figure 3). We approximated the distribution by fitting an exponential function of the form $P(x) = Ae^{-\lambda x}$ in a least squares sense. We then used $P(x)$ to draw sample values during the missing samples restoration phase.

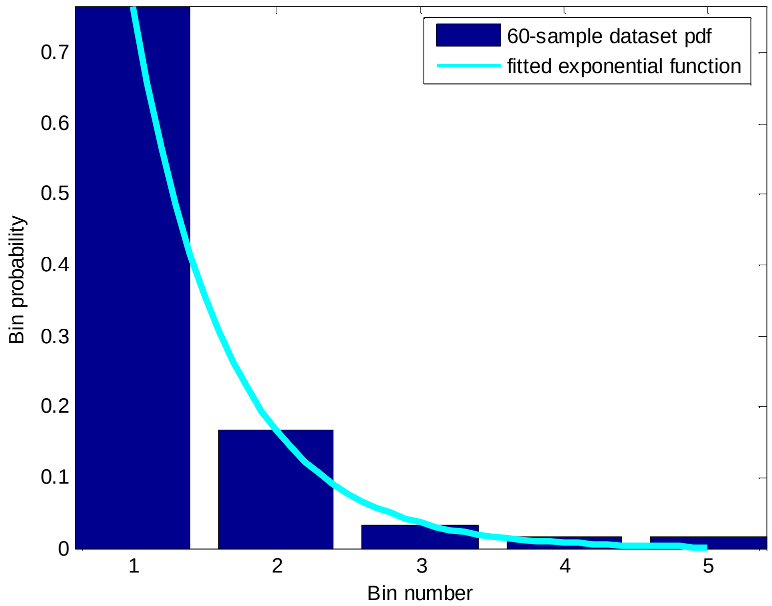


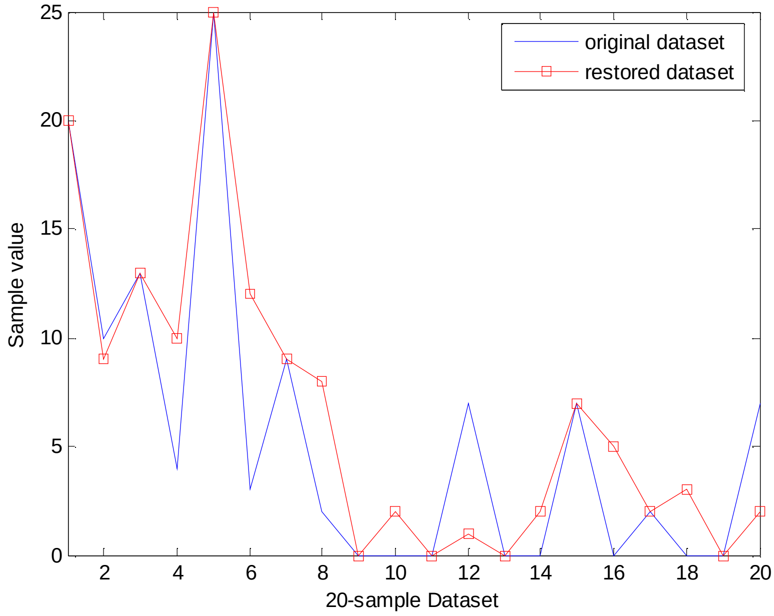
Figure 3 - Probability density function of the original 60-sample dataset and the fitted exponential function.

To restore the missing samples we have combined the linear gradient estimation function $g(x)$ with the global population distribution function $P(x)$.

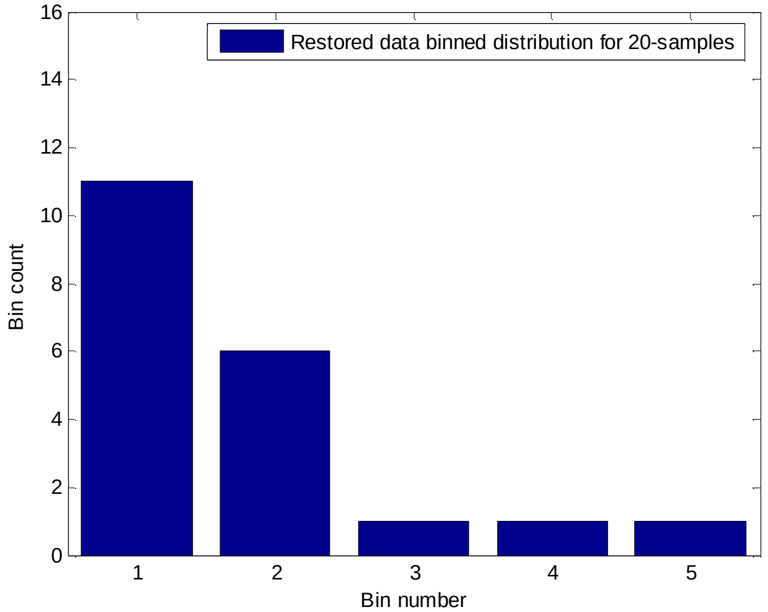
Indeed, suppose we wish to estimate the sample value at x. We locate previous known sample, at x-1 (as a reminder, we have removed every other sample from the 20-sample dataset to derive the 10-sample dataset). We then use the gradient estimation function $g(x)$ to get an approximation of the gradient for the value at x-1. This gradient gives us a "locality range" in which we expect the estimated value $f(x)$ at x to lie in, i.e. we assume that the estimated value will be larger or smaller than the value at x-1 by a small amount (neighborhood effect).

We then draw a random number $f(x)$ from the $P(x)$ distribution that is within the above range and use that as the sample value estimation. The restored sample values will thus follow the distribution of the original dataset as well as follow the localized restrictions of the dataset.

Figure 4a demonstrates the approximation achieved by this method. According to the correlation coefficient ρ which is calculated as 0.85, the restored dataset is close to the original 20 sample dataset. Furthermore, comparing the binned distributions of the restored and original 20-sample dataset (figures 1c and 4b) using a chi-squared test provide a χ^2 value of 0.28 which, for 4 degrees of freedom is much smaller than the chi-squared threshold of 9.448 for a 0.05 significance level.



(a)



(b)

Figure 4 - The original and approximated 20-sample dataset values (a) and the binned distribution function of the restored dataset (b)

5 Conclusions

The proposed formalism incorporates local and global statistical feature of arbitrary sampling data in order to achieve more efficient interpolation procedure. The formalism lies to the assumption (and this introduces a limitation of the present work) that there is a critical sample size for which statistical properties coincides with the statistical properties of the entire system under examination.

Theoretical estimation of the parameters enters in the proposed formalism is a part of a future work. Indeed, the estimation of how dense the original sampling set should be, in order to extract the gradient behaviour of sample data, is missing. More over, the role of spatial heterogeneity in estimating the range over which interpolation remains a robust procedure should be clarified.

It is noted that while the application given in this work refers to ecological spatial sampling data, the formalism may extend to time series data and interpolation or extrapolation problems.

6 Acknowledgements

Financial support of the European Regional Development Fund (ERDF) and of the Region of Ionia Nisia under the Project “Modeling self-organization dynamics and suppression of dacus population in real Ecosystem of Saint George Municipality in Corfu” (IN_9) is gratefully acknowledged.

References

Ahnert, K. and Abel, M. (2007) Numerical differentiation of experimental data: local versus global methods, *Submitted to Computer Physics Communications.*

Avlonitis, M., Mangkos, E., Stefanidakis M. And Chrissikopoulos V (2007), A spatial stochastic model for worm propagation: scale effects, Journal of Computer Virology, 3, pp. 87-92.

Bueno, A.M and Jones, O. (2002) Alternative methods for controlling the olive fly, Bactrocera oleae, involving semiochemicals, IOBC wprs Bulletin, 25.

Burnett DS (1987) Finite element analysis, from concepts to applications, Addison-Wesley

Fletcher, B. S. and Kapatos, E. T. (1981) Dispersal of the olive fly, Dacus oleae, during the summer period on Corfu. Ent. Exp. Appl., 29, 1-8. Ned. Entomol. Ver. Amsterdam.

Kapatos, E.T and Fletcher, B.S. (1983) Establishment of economic injury levels for olive infestation by Dacus oleae in Corfu. Ent. Hell. 1, pp. 37-45

Kapatos, E.T, Mc Fadden, M.W. and Pappas, S. (1977) Sampling techniques and prepation of partial life tables for the olive fly, Dacus oleae, (Diptera, Tephritidae) in Corfu. Ecol. Ent., 2, pp. 193-196.

Peters, D.P.C. and Herrick J.E. (2004) Strategies for ecological extrapolation, OIKOS, 106(3), pp. 627-636.

Piegl L (1991) On NURBS: a survey. IEEE Comput Graph Appl 11(1), pp. 55–71

Shepard, D. (1968). A two-dimensional interpolation function for irregularly spaced data. In: Proceedings of the 1968 ACM National Conference, pp. 517–524.