Entomological data collection and unreal assumptions

L.V. Nedorezov

University of Nova Gorica, Vipavska Cesta 13, Nova Gorica SI-5000, Slovenia
e-mail: lev.nedorezov@ung.si

Abstract

In current publication a stochastic model of individual migrations within the limits of finite domain, and process of data collection are considered. It is assumed that for every set of data collection population size is constant, local population size is fixed in 100 different points on plane, and time interval between two nearest moments of population size measurements is rather big. It is also assumed that population size scale isn't homogeneous, and there exists an interval of optimal values of population size: all individuals try to migrate (with biggest probability) to the respective part of plane. For artificial datasets hypotheses about correspondence between real average of population density and estimations obtained for different sample sizes were checked. Hypotheses about correspondence of observed samples to Normal distribution when sample size $N = 7,8,...,100$, were also checked with Kolmogorov-Smirnov test.

Key words: stochastic mathematical model, migrations, data collection

Introduction

Estimations of mathematical model parameters of population dynamics using real time series are among of the main directions in creating of forecasts of population size changing in time, in finding of optimal methods of population size management etc. In many cases the process of model parameters estimation starts with the words "Let's consider a time series of population density (or population size) changing in time…" $\{x_k^*\}$, $k = 0,...,N$, where $N+1$ is a sample size. This investigation must also start with the following key words: "Let's consider the following mathematical model with unknown parameters which describes the population density changing in time":

$$\frac{dx}{dt} = F(t, x, \alpha). \quad (1)$$

In equation (1) $x(t)$ is a population density at time t; α is a vector of unknown model parameters; F is non-linear function which satisfies to a set of known limits (Brauer, Castillo-Chavez, 2001; Nedorezov, Utyupin, 2011; McCallum, 2000; Kendall et al., 1999, 2005; Turchin, 2003 and others). Model (1) may be of other type, for example, it can be a system of recurrence equations, but the problem will be the same: for existing time series $\{x_k^*\}$ values of model (1) parameters α must be determined. It is important to note, that initial value of population density $x_0 = x(0)$ is unknown parameter too, and it must be determined using existing datasets. Even in the case when we have experimental datasets, initial value x_0 is known amount and doesn't need to be determined (Nedorezov, 2011, 2012 a, b).

It is possible to point out a lot of various approaches to the problem of estimation of model parameters (see, for example, Pawitan, 2001; McCallum, 2000; Wood, 2001a, b). Least squares method is one of the basic methods and it is widely used in practice (Hudson, 1970; Demidenko, 1981; Gubarev, 1985; Nedorezov, Sadykova, 2005, 2008, 2010). The use of least squares method assumes that we have to find the global minimum for squared deviations between theoretical (model) values (which must be obtained with model (1)) and empirical dataset $\{x_k^*\}$. For example, we have to find the global minimum for the following functional form:

$$Q(\alpha, x_0) = \sum_{k=0}^{N} (x_k^* - x(t_k, \alpha, x_0))^2 . \qquad (2)$$

In (2) $x(t_k, \alpha, x_0)$ is the respective theoretical value (for *global fitting* it is a solution of equation (1)) which is obtained with equation (1) for concrete values of model parameters α and initial value of population density x_0.

There are some basic requirements which are assumed to be realized for suitable models. In particular, deviations between theoretical and empirical values must correspond to Normal distribution with zero average (Hudson, 1970; McCallum, 2000; Wood, 2001a, b; Nedorezov, 2011, 2012 a, b). In this case the natural questions arise: what is the base for the assumption that deviations between theoretical and empirical values must have Normal distribution? What is the base for the assumption that all deviations have one and the same Normal distribution? But the answers are rather obvious: there are no real bases which are determined by the biology of investigated object, for both these assumptions.

Requirement about the equivalence of arithmetic average to zero is correct and obvious – in a set of measurements the systematic errors cannot be observed. It is also obvious that distribution of deviations must be a symmetric function with respect to origin, and realization of bigger deviation must be observed with smaller probability. From these two obvious requirements we can't conclude that distribution of deviations is Normal. It is important to note that assumption about normality of deviations is in contradiction with sense: for example, if we estimate the weight of larva in milligrams we cannot have a mistake in several tons with positive probability in principle. We cannot also observe a negative value of weight with positive probability.

In current publication the following important problems are analyzed. First of all, taking into account that "real" local population size is known in model, it is possible to estimate the probability that "real" population density belongs or doesn't belong to confidence interval which is determined with standard methodology (with the help of Student's distribution). These calculations were provided for different values of "real" population density and different sample sizes $N = 7,8,...,100$. It was obtained that for small sample size and small population density number of results when "real" population density doesn't belong to confidence interval can be up to 50%. Increase of sample size leads to decreasing of number of cases when "real" population density doesn't belong to confidence interval; in all modeled situations this number wasn't equal to zero.

The second, for different values of "real" population density and different sample sizes $N = 7,8,...,100$ analysis of correspondence of artificial samples to Normal distribution was provided with Kolmogorov-Smirnov test. It was obtained that for small samples number of cases when hypothesis about Normality must be rejected, can be very small. At the same time the following situations were observed in the model: increasing of sample size led to increase the number of cases when hypothesis about Normality must be rejected (up to 100%).

Taking also into account that dependence of sample variation on population density may have a non-linear and non-monotonic character (Nedorezov, 2012 b, c, 2013), we can conclude that even for very simple situations there are no reasons for assumptions about Normality of initial samples, no reasons for assumptions about equivalence of sample variations for different values of population densities, and probability that real average of population size belongs to confidence interval (when limits of interval are obtained with standard method of parametric statistics) can be very small.

Description of the model

Let N be a total population size. We'll assume that $N = const$ during the time of providing of computer experiments (for estimation of population density). And let Z_{nm}^2 be an integer rectangular lattice on the plane R^2:

$$Z_{nm}^2 = \{(i, j) : 1 \le i \le n, 1 \le j \le m\}.$$

Additionally we'll assume that local population size is determined in knots (i, j) of the lattice Z_{nm}^2. Denote it as $x_{ij}(t)$ for $(i, j) \in Z_{nm}^2$ at time moment t. For all time moments t, $t = 0,1,2...$, the following relation is truthful:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}(t) = N.$$

It means that there are no migrations outside the domain Z_{nm}^2.

Definition. We'll call two knots (i_1, j_1), $(i_2, j_2) \in Z_{nm}^2$ as neighboring knots if and only if the following relation is truthful:

$$|i_1 - i_2| + |j_1 - j_2| = 1.$$

Within the framework of model it will be assumed that migration processes from the knot (i, j) can be observed to neighboring knots only. About the behavior of migrants we'll assume that for all values of local population sizes there is the quota δ, $\delta = const > 0$, $\delta < 1$, of individuals which migrate to neighboring knots with equal probabilities (behavior of these individuals doesn't depend on current situations in Z_{nm}^2). Let's also assume that probabilities of migration of all other individuals to neighboring knots (the quota of these individuals is equal to $1 - \delta$) depend on a distribution of individuals in neighboring knots of the lattice Z_{nm}^2.

We'll also assume that all knots of the lattice Z_{nm}^2 are divided with respect to local population size on to three qualitatively different types. Denote as D_1 and D_2, $D_1 < D_2$, two critical levels which determine the optimal interval of population size; respectively, we'll assume that if in knot (i, j) the population size $x_{ij}(t)$ satisfies to the following inequalities $D_1 \le x_{ij}(t) \le D_2$ from this knot we can observe stochastic migrants only (total number of stochastic migrants is about $\delta x_{ij}(t)$). In this situation the number of non-migrated individuals is about $(1 - \delta) x_{ij}(t)$.

If the following inequality is truthful $x_{ij}(t) < D_1$, all individuals try to leave this knot (they migrate to neighboring knots). In the situation when we have very high local population size, $x_{ij}(t) > D_2$, the considering system is out of the optimal zone, and in this situation we'll also assume that individuals try to leave this knot. But for this situation we'll have one more probability – the probability for staying in this knot.

Denote as q_j weights which correspond to level of attractiveness of knots for migrants, $q_j = const > 0$. We'll assume that attractiveness of knot where local population size is less than D_1, is equal to one, $q_1 = 1$. Attractiveness of knot q_2 where local population size belongs to optimal interval, must be bigger than one, $q_2 = const > 1$. We'll also assume that attractiveness of knot q_3 where local population size is bigger than D_2, $x_{ij}(t) > D_2$, has minimal value, $q_3 < q_1$ (it is possible to point out species and situations when this assumption isn't truthful, and migrants move to domain where population density is extremely high; Isaev et al., 1984, 2001).

Let $q_{ij}(t)$ be an attractiveness of knot (i, j) at moment t (it is obvious, that depending on local population size $x_{ij}(t)$ this amount $q_{ij}(t)$ will be equal to q_1, q_2 or q_3). Let's assume that for every individual probability to migrate to one of nearest knots is proportional to attractiveness of this knot and inversely to sum of weights of all neighbour knots. Thus, when $x_{ij}(t) < D_1$ and all individuals try to escape out this knot, and knot (i, j) doesn't belong to the boundary of lattice Z^2_{nm}, we'll assume that probability p_{i-1j} of migration of every individual to knot $(i-1, j)$ is defined by the following expression:

$$p_{i-1j} = \frac{q_{i-1j}(t)}{q_{i-1j}(t) + q_{i+1j}(t) + q_{ij-1}(t) + q_{ij+1}(t)}. \quad (3)$$

All other three probabilities are of the same type. For the case $x_{ij}(t) > D_2$ probability for individual to stay at the same knot will be defined by the following expression:

$$p_{ij} = \frac{q_{ij}(t)}{q_{ij}(t) + q_{i-1j}(t) + q_{i+1j}(t) + q_{ij-1}(t) + q_{ij+1}(t)}. \quad (4)$$

Note, that migration flow (which is determined by the expressions (3) and (4)) increases monotonously with increase of coefficient of attractiveness. In expression (4) $q_{ij}(t) = q_2$.

Results of modeling

For the computer modeling of migration processes it was assumed that total population size N is constant; thus, theoretical population density μ was known and equal to $\mu = N/nm$. Below we present results of modeling for the quadratic lattice Z_{nm}^2 with $m = n = 100$. Initial population state was modeled in the following manner: every individual with equal probabilities could appear in every knot of the lattice. After determination of initial situation the process of individual's migrations was started (with respect to formulas (3)-(4)). During the time T (number of time steps; for providing calculations it was assumed that $T = 1000$) was run free. It is important moment because we have to have on the lattice the situation which is determined by the population migration process only, and doesn't depend on the initial state of population.

After that the process of data collection was started: in 100 different stochastic points of the lattice Z_{nm}^2 the local population size was fixed (it looks like 100 casts of the frame). After that model run free during the next T time steps; after that we had the same process of data collection and so on. This procedure was repeated 100 times for every fixed population size N. Let's consider in details the situation with $D_1 = 10$, $D_2 = 30$, $\delta = 0.1$. For realization of stochastic process on the computer it was assumed that weight $q_1 = 1$ if the local population size is less than D_1; weight $q_2 = 10$ if the local population size belong to optimal zone $[D_1, D_2]$; weight $q_3 = 0.2$ if local population size is greater than D_2.

When population density is small (fig. 1a, density is equal to 0.1) real population density is out off confidence interval in 49% of all cases (number of measurements is equal to 7). Increase of number of measurements leads to decrease of number of errors (fig. 1a) but it may have non-monotonic character and can be big (21%) if number of trials is equal to 52. But it looks like unrealistic situation in entomological investigations. Increase of population density in 10 times (fig. 1b) leads to decrease of number of errors – maximum (10%) of errors is observed when number of trials is equal to 7. Big number of errors (8%) can be observed for unrealistic big number of trials – 34 (fig. 1b). Further increasing of population density (fig. 1c) can lead to increase of maximum of errors we may have in estimation of population density: when number of trials is equal to 7 number of errors is equal to 34%. About 14% of errors we may have for 35 trials (fig. 1c).

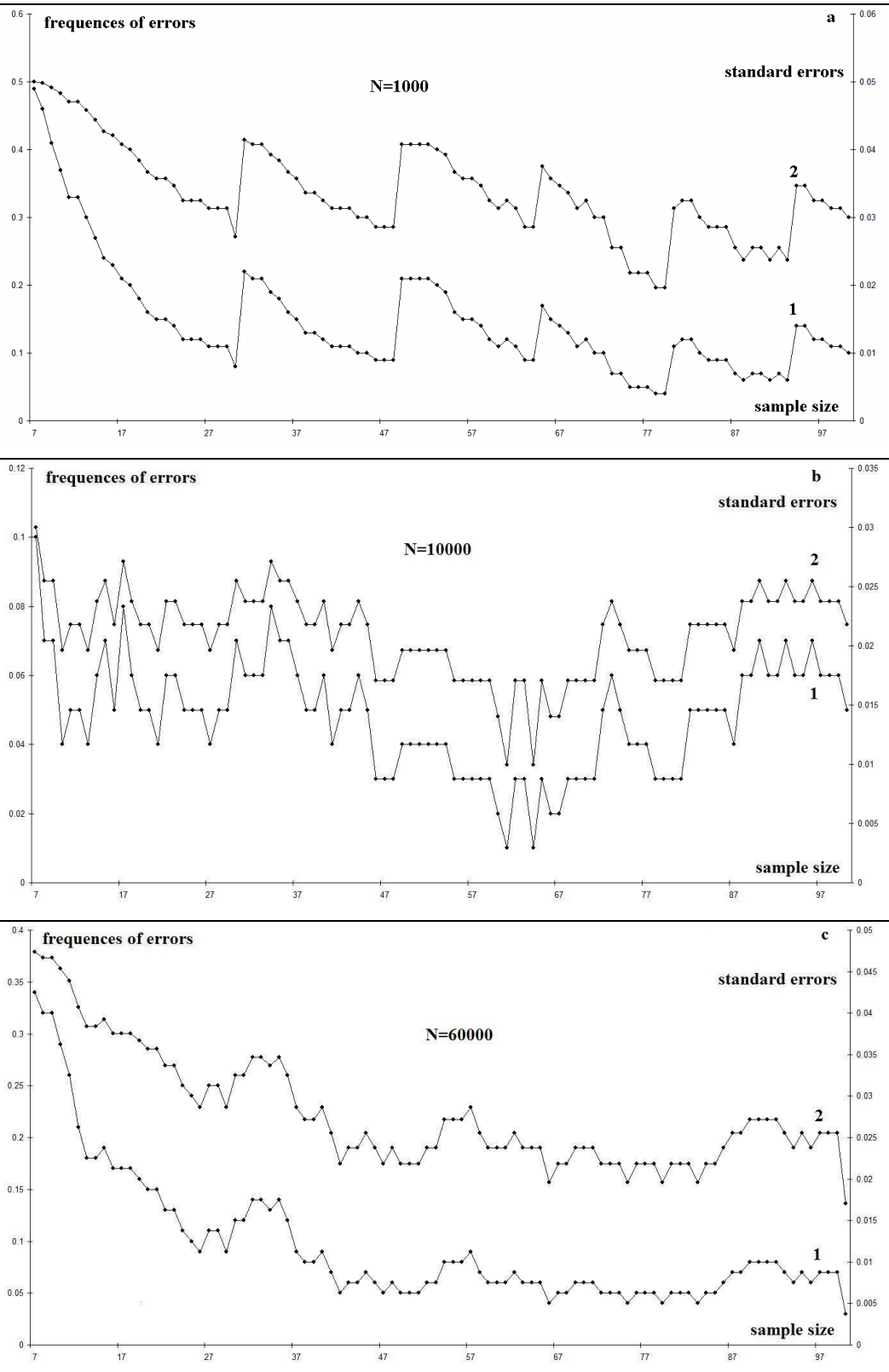

Fig. 1 Dependence of errors (frequencies of cases when real average doesn't belong to confidence interval) on sample size, and standard errors of these frequencies. a – population density is equal to 0.1; b - population density is equal to 1.0; c - population density is equal to 6.0.

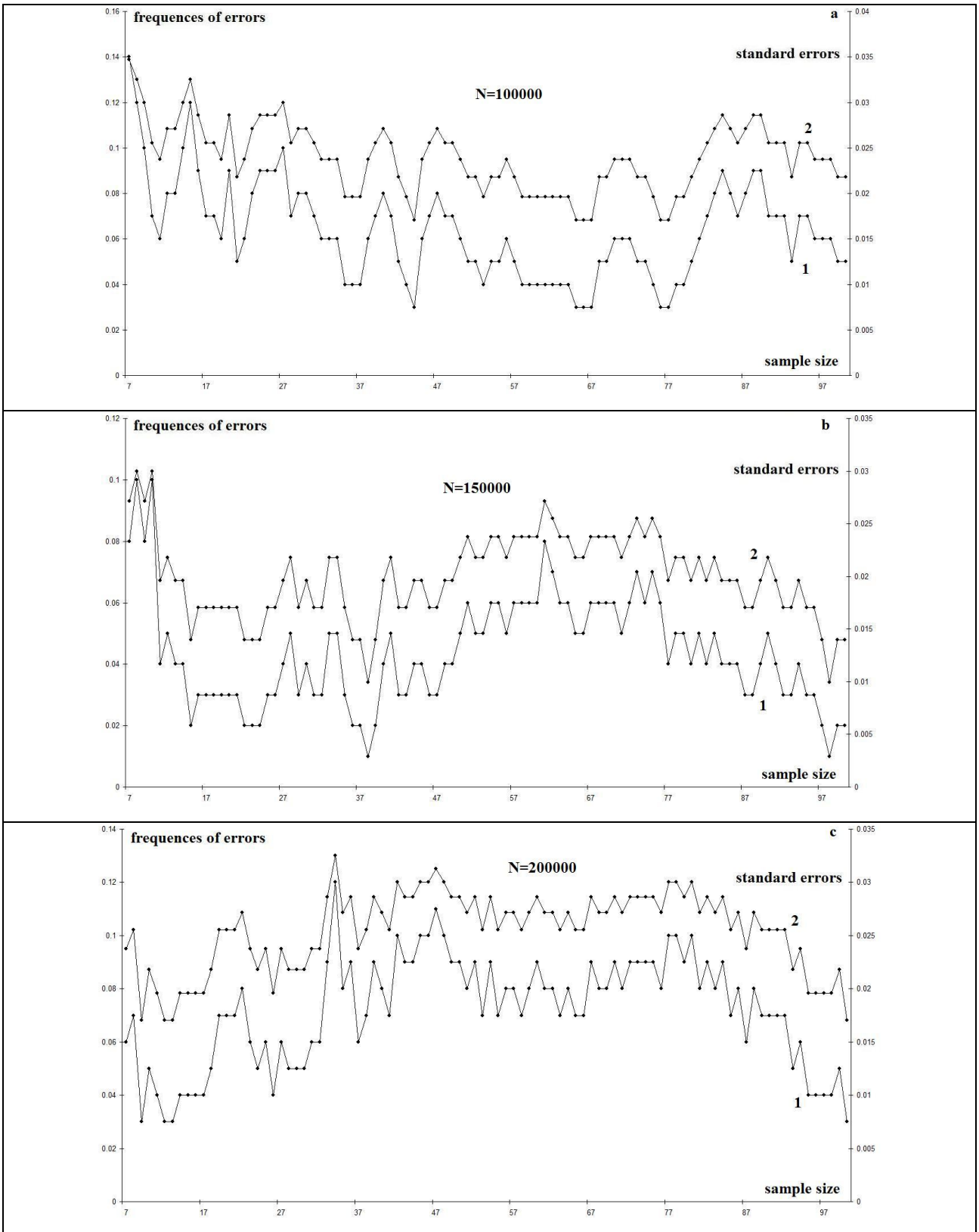

Fig. 2 Dependence of errors (frequencies of cases when real average doesn't belong to confidence interval) on sample size, and standard errors of these frequencies. a – population density is equal to 10.0; b - population density is equal to 15.0; c - population density is equal to 20.0.

On fig. 2 there are the pictures which show changing of errors on sample size (number of trials for the estimation of average of population density) when real population density is big. In such situations we have strong influence of existence of optimal zone $[D_1, D_2]$ on distribution of individuals, their behavior, and, respectively on collected results. In first case (fig. 2a) we can see rather small variation of frequency of errors (from 3% to 14%); in other situations (fig. 2 b, c) we can observe non-monotonic behavior of errors with increase of sample size. In particular (fig. 2c) maximum of errors (12%) is observed for sample size 28.

Thus, obtained results show us that when population size is sufficient small provided measurements and standard statistical estimations of averages may correspond to nothing. If number of measurement is equal to 7 number of errors – number of cases when real average doesn't belong to confidence interval – is equal to 49%. If number of trials is less then 7 number of errors will increase. And the second, well-known idea in statistics that increase of sample size must lead to obtaining of better results (in particular, in estimation of average) isn't truthful for considering situations. Number of errors in estimation of averages may have non-linear character with respect to number of trials, and can increase with increase of trials (fig. 2).

In our publications (Nedorezov, 2012 b, c) it was obtained that deviations between "real" population density and results of computer experiments may have non-monotonic character with respect to (fixed) population density. This complicated behavior can be explained as a result of influence of non-homogenous structure of distribution of individuals on a plane and existence of optimal zones of individual's concentration. Within the framework of considering model we analyzed also dependence of sample variations on population density. Results of analysis of computer experiments are presented on fig. 3.

For every fixed population density variations were estimated with formulas:

$$s_1^2\left(\frac{N}{nm}\right) = \frac{1}{9999}\sum_{k=1}^{10000}\left(x_k - \frac{N}{nm}\right)^2, \tag{5}$$

$$s_2^2\left(\frac{N}{nm}\right) = \frac{1}{4999}\sum_{k=1}^{5000}\left(\frac{x_{2k-1}+x_{2k}}{2} - \frac{N}{nm}\right)^2. \tag{6}$$

In (5) and (6) x_k are the "empirical" values, N/nm is "real" population density. On the graphics of these functions it is possible to point out two critical points where behavior of functions changes cordially (fig. 3). First point is near amount 1.7: near this point influence of optimal zones becomes rather strong and it leads to bigger heterogeneity in distribution of individuals on

the plane. In a result of this influence increasing of functions (5) and (6) is much faster than it can be observed before.

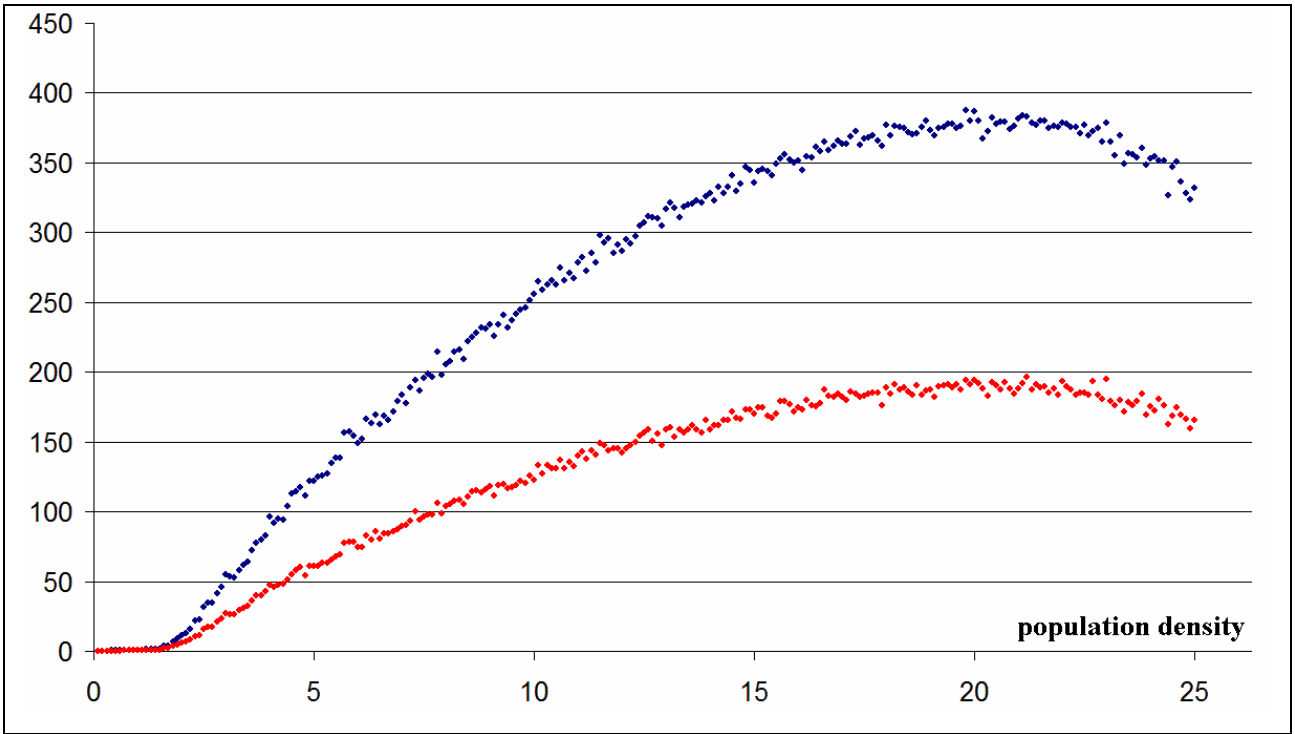

Fig. 3 Changing of sample variations (5) (blue points) and (6) (red points) at increase of population density.

The second critical point is in between 20 and 20.2 (fig. 3): influence of non-optimal zones $\{x > D_2\}$ becomes strong and it leads to decreasing of variations (5) and (6). This non-monotonic behavior of sample variations was obtained within the limits of very simple mathematical model. Taking into account that for natural populations changing of behavior of individuals at changing of population density is usually unknown and can be very complicated, changing of sample variation can be very difficult. It allows us to conclude that in analysis of empirical time series of natural population problem of heteroscedasticity will never be solved.

The next important question is a question about Normality of initial samples. For testing of Normality of samples for different amounts of population density the well-known Kolmogorov – Smirnov criteria was used. On figure 4 there are several curves which show the dependence of number of cases when we have to reject Null hypothesis about Normality of samples on sample size for fixed values of population densities. As we can see on this fig. 4, if population size is very small (curve with $N = 1000$) for sample size 7 Null hypothesis must be rejected in 49% of

all cases. Number of these cases increases very fast and when sample size is equal to 12 Null hypothesis must be rejected in 100% cases (with 5% significance level).

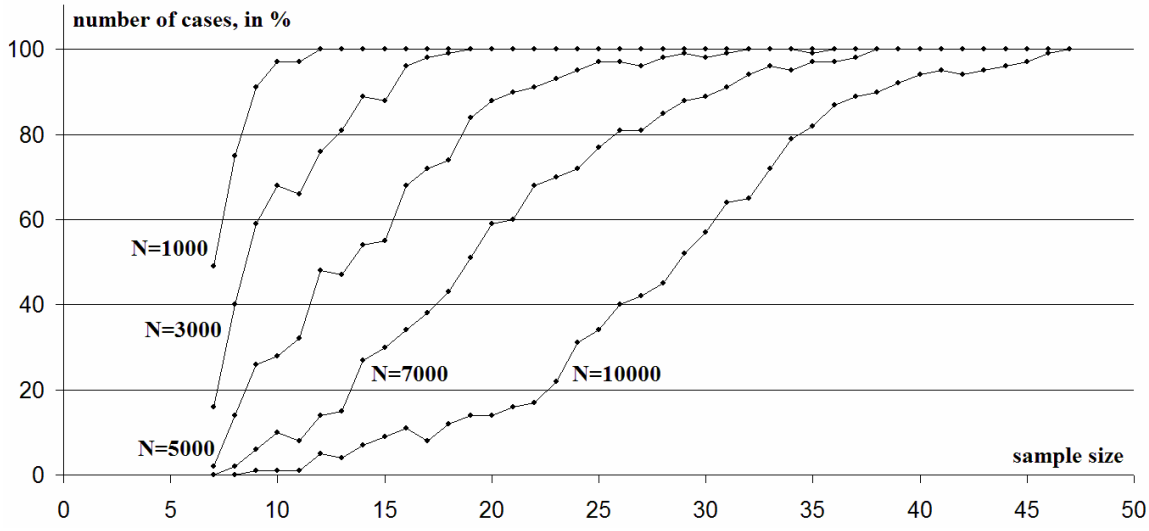

Fig. 4 Changing of number of cases when hypothesis about Normality of sample must be rejected on sample size for different values of population density.

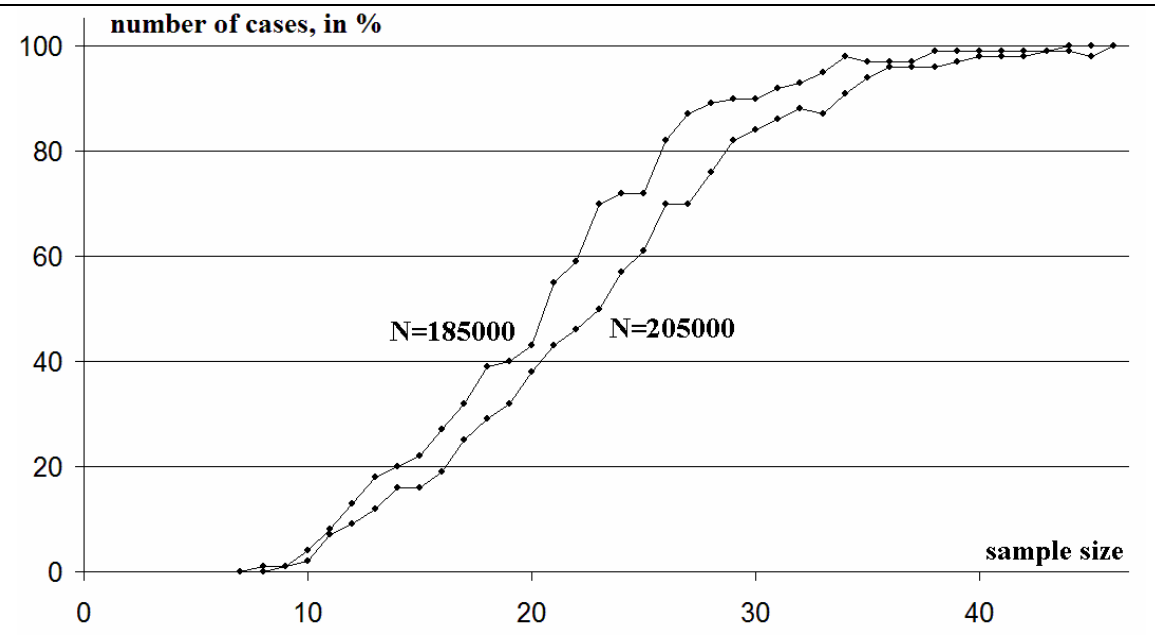

Fig. 5 Changing of number of cases when hypothesis about Normality of sample must be rejected on sample size for different values of population density.

When population size is bigger than 0.1, we can observe the similar behavior of number of cases: in all situations there exists critical number of sample size when number of cases is equal to 100%. When density is equal to 1.0 this critical value is 47 (fig. 4). The similar situation we

can observe for bigger values of population density (fig. 5). Critical points for cases when $N = 185000$ and $N = 205000$ are equal to 38 and 40 respectively. Thus, analysis of simple mathematical model allows us concluding that for sufficient big sample sizes and for various population densities we have to reject hypothesis about Normality of initial samples we can get with method of "casting of the frame" (and also for other closed methods of biological data collection). For natural populations which have more complicated mechanisms of interaction between individuals, we cannot assume that situation with initial samples will be better.

Conclusion

Considered stochastic model of the process of data collection looks like well-known method of "casting of the frame". Within the framework of model it was assumed that migration processes of individuals depends on the conditions of knots of lattice. Condition of the knot had been determined with respect to interval of population size. Knot had a highest weight (for attraction of individuals) if population size was in optimal interval. Knot had smaller weight if population size was less than for optimal zone. And, finally, knot had smallest weight if population size was bigger than for optimal zone.

Computer experiments with model show that when population size is sufficient small standard statistical estimations of averages may correspond to nothing. If number of measurement is equal to 7 number of errors – number of cases when real average doesn't belong to confidence interval – is equal to 49% (fig. 1a). It is obvious, if number of trials is less then 7 number of errors increases. Computer experiments allow also concluding that well-known idea in statistics that increase of sample size must lead to obtaining better results (in particular, in estimation of average) isn't truthful for considered situations. Number of errors in estimation of averages may have non-linear character with respect to number of trials, and it can increase with increase of trials (fig. 2). Computer experiments allowed also concluding that hypothesis of Normality of collected samples must be rejected in 100% cases with 5% significance level for sufficient big sample sizes and for various values of population density.

References

Brauer F., Castillo-Chavez C. 2001. Mathematical Models in Population Biology and Epidemiology. N.Y.: Springer-Verlag

Demidenko E.Z. 1981. Linear and non-linear regression. Moscow: Finance and statistics.

Gubarev V.V. 1985. Algorithms of statistical measurements. Moscow: Energoatomizdat.

Hudson D.J. 1970. Statistics for physicists. Lecture notes on elementary statistics and probability. Moscow: Mir.

Isaev A.S., Khlebopros R.G., Nedorezov L.V., Kondakov Yu.P., Kiselev V.V. 1984. Forest Insect Population Dynamics. Novosibirsk: Nauka.

Isaev A.S., Khlebopros R.G., Nedorezov L.V., Kondakov Yu.P., Kiselev V.V., Soukhovolsky V.G. 2001. Population Dynamics of Forest Insects. Moscow: Nauka.

Kendall B.E., Briggs C.J., Murdoch W.W., Turchin P., Ellner S.P., McCauley E., Nisbet R.M., Wood S.N. 1999. Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. Ecology, 80(6): 1789-1805.

Kendall B.E., Ellner S.P., McCauley E., Wood S.N., Briggs C.J., Murdoch W.W., Turchin P. 2005. Population cycles in the pine looper moth: dynamical tests of mechanistic hypotheses. Ecological Monographs, 75(2): 259-276.

McCallum H. 2000. Population parameters estimation for ecological models. Brisbane: Blackwell Sciences.

Nedorezov L.V. 2011. Analysis of some experimental time series by Gause: Application of simple mathematical models. Computational Ecology and Software, 1(1): 25-36.

Nedorezov L.V. 2012 a. Gause' Experiments vs. Mathematical Models// Population Dynamics: Analysis, Modelling, Forecast 1(1): 47-58.

Nedorezov L.V. 2012 b. Chaos and Order in Population Dynamics: Modeling, Analysis, Forecast. LAP Lambert Academic Publishing.

Nedorezov L.V. 2012 c. Modeling and analysis of some methods of entomological data collection// Computational Ecology and Software 2(2): 83-95

Nedorezov L.V. 2013. About some background problems for ecological modelling// Population Dynamics: Analysis, Modelling, Forecast 2(1): 23-37

Nedorezov L.V., Sadykova D.L. 2005. A contribution to the problem of selecting a mathematical model of population dynamics with particular reference to the green oak tortrix. Euro-Asian Entomological Journal, 4(4): 263-272.

Nedorezov L.V., Sadykova D.L. 2008. Green oak leaf roller moth dynamics: An application of discrete time mathematical models. Ecological Modelling, 212: 162-170.

Nedorezov L.V., Sadykova D.L. 2010. Analysis of population time series using discrete dynamic models (on an example of green oak leaf roller). Lesovedenie, 2: 14-26.

Nedorezov L.V., Utyupin Yu.V. 2011. Continuous-Discrete Models of Population Dynamics: An Analytical Overview. Novosibirsk: State Public Scientific-Technical Library SB RAS.

Pawitan Y. 2001. In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford: Clarendon Press.

Turchin P. 2003. Complex Population Dynamics: A Theoretical/Empirical Synthesis. Princeton: Princeton University Press.

Wood S.N. 2001a. Minimizing model fitting objectives that contain spurious local minima by bootstrap restarting. Biometrics, 57: 240-244.

Wood S.N. 2001b. Partially specified ecological models. Ecological Monographs, 71: 1-25.