Nearest Neighbour Distance Statistics Estimation

Jiřina, Marcel
2002

cAk

CENTRUM APLIKOVANÉ KYBERNETIKY

České vysoké učení technické v Praze - fakulta elektrotechnická

Nearest Neighbour Distance Statistics Estimation

Technical report

Marcel Jiřina

www@c-a-k.cz

2002

Institute of Computer Science
Academy of Sciences of the Czech Republic

Nearest Neighbour Distance Statistics Estimation

Marcel Jiřina

Technical Report No. 878

October 2002

Abstract

In this report we show that the square of nearest neighbour distance (NND) can be approximated by Erlang distribution for large number of service lines and n-1 used lines, which includes case of n = 1, too.

Keywords:
nearest neighbour distance, Erlang distribution

Introduction

To compute the nearest neighbour distance (NND) is rather simple but tiresome task especially for large *n*, the space dimension. The other task is what the statistical distribution of the nearest neighbour distance. For *n* = 1 it is simply exponential distribution. In this report we show that the square of NND can be approximated by Erlang distribution for large number of service lines and *n*-1 used lines, which includes case of *n* = 1, too.

Erlang distribution

Erlang distribution [1], [2] was derived by Erlang in 1902 for description and evaluation of traffic in telephone switchboards. The simplest task considers N service trunks (more generally service lines) available. The telephone calls (tasks) come with intensity of x Erlangs, i.e. in average x tasks in time unit; x is positive real number. It is supposed that arrivals and terminations are independent events and then one can use Poisson distribution for call arrivals. It is also assumed that a statistical equilibrium exists that is that the number of calls in progress remains the same, i.e. the number of new call arrival equals to the number of terminations. If there is no free trunk the call is cleared; there is no queue to await a free path. In given time r calls are in progress, r=0, 1, ... N.

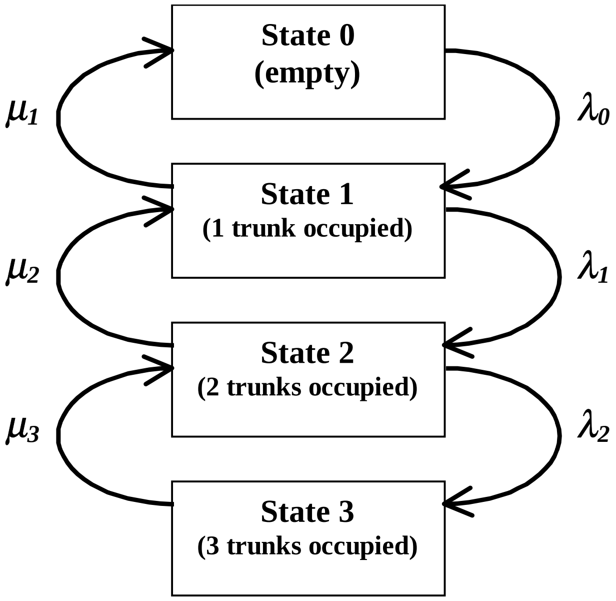

Fig.1. State model of three trunk system.

The system considered could be represented by state diagram [1], see Fig. 1. There is model of *N* = 3-trunk system that has four states corresponding to 0 - 3 trunks occupied. At each state we have a probability of a call arrival λ_r and the probability of a call termination μ_r . Since we assume we are in equilibrium the numbers of call arrivals must equal the number of call terminations and so the probability of a call arrival must equal to the probability of call termination.

The basic formula for probability density of Erlang system of total *N* trunks, *r* trunks occupied and with load of *x* Erlangs is

$$p(N,r,x) = \frac{\frac{x^r}{r!}}{\sum_i^N \frac{x^i}{i!}} \quad .$$

Sometimes so called scale factor α is introduced so that $x = \alpha y$. For unlimited sources, i.e. number of trunks available $N = \infty$, it holds

$$p(r,x) = \frac{x^r}{r!} e^{=x} \quad .$$

In this case the distribution function is

$$F(r,x) = 1 - e^{-x} \sum_{j=0}^{r-1} \frac{x^j}{j!} \quad .$$

Note that Erlang distribution is a particular form of gamma distribution [2].
Numerical characteristics of Erlang distribution are as follows [2] (we suppose $x = \alpha y$)
Expected value $E(x) = r/\alpha$, variance $V(x) = r/\alpha^2$, skewness $g_1 = 2/\sqrt{r}$, kurtosis $g_2 = 3 + 6/r$.
Moreover it holds [2]:
Lemma 1: The sum of r independent exponentially distributed variates with parameter α is equal to an Erlang (gamma) distributed variate with parameters α and r, i.e. let $Y_i \approx \exp(\alpha)$ and $X = \sum_{i=1}^{r} Y_i$ then $X \approx Erl(r,\alpha)$.

The nearest neighbour distance

There are some tasks that use either the nearest neighbour or its distance from particular given point in *En*. One application is the distribution density estimation by the nearest neighbour or 5th nearest neighbour approach used in Bayes classifiers [3], [4]. We limit ourselves the distance of the nearest neighbour, to Euclidean space only, and to homogenous distribution of points randomly placed in *En*.

E1 case

In this simplest case let in the interval of length l_0 be placed n_0 points randomly but with homogenous distribution. The average distance of two successive points is simply $\bar{d}_p = l_0 / n_0$ and it is known that these neighbour intervals have exponential distribution with parameter $\bar{d}_p$. The NND has also exponential distribution but with parameter $\bar{d}_1$, $\bar{d}_1 = \bar{d}_p / 2$.
It can be easily found that of all neighbour intervals 1/3 is not „used“ for NND, 1/3 is used twice as for both neighbours the other end point of particular neighbour interval is just the nearest neighbour, and 1/3 neighbour intervals are used once.

En case

Following idea of Lemma 1 we can take the *n*-dimensional Euclidean space as composed from *n* one-dimensional spaces. Considering randomly distributed points in *En* from the homogenous distribution, the marginal one-dimensional distributions are homogenous. Then for each of them the distribution of the nearest neighbour distance has exponential distribution. From the Lemma 1 then holds that nearest neighbour distance distribution in *En* is Erlang distribution with unlimited sources E(r,α), where r = N-1.
Another possible point of view to *En* is that to each dimension in this space let us assign one state in the sense of Erlang mass service system [1], [2], see Fig 1.

Numerical experiments

In all numerical experiments were used approximately 32000 or 1000 points in unit cubes in E1, E2, E3, spaces. For E10 3000 points was used.

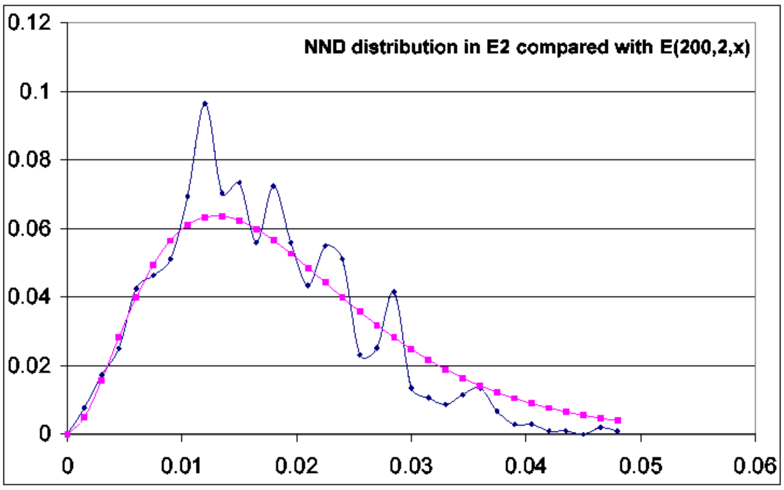

Fig. HD2^1Erl.ps. Comparison of Erlang distribution with NND distribution in E2.

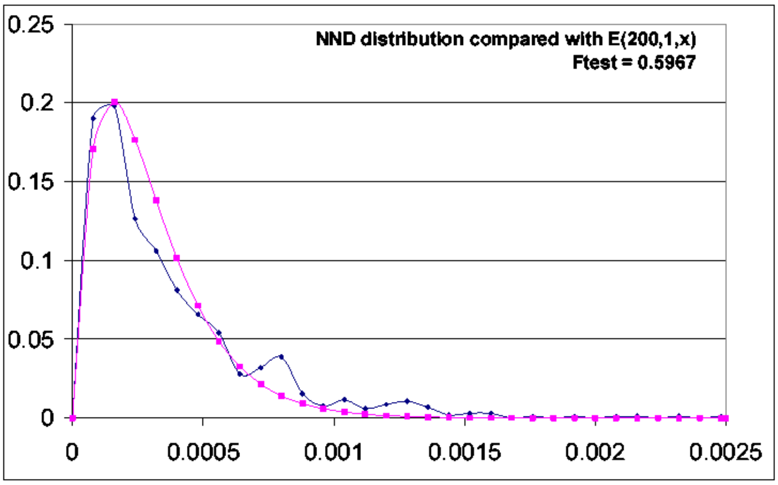

Fig. E200-1-x.ps. Comparison of Erlang distribution with NND^2 distribution in E2.

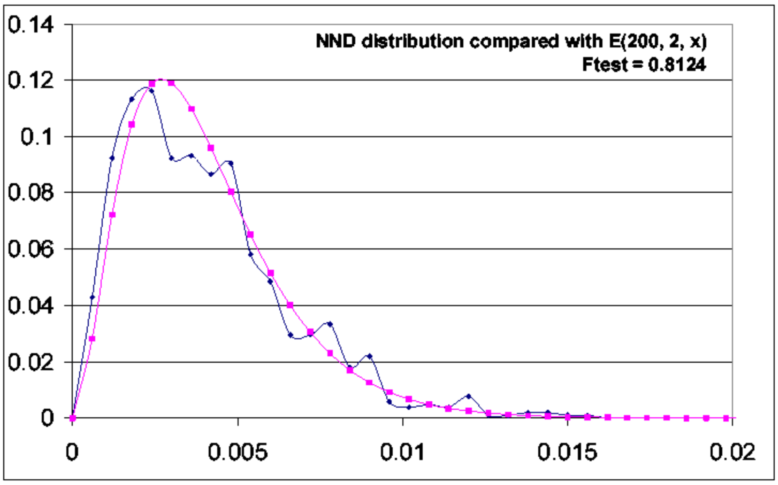

Fig. E200-2-x.ps. Comparison of Erlang distribution with NND^2 distribution in E3.

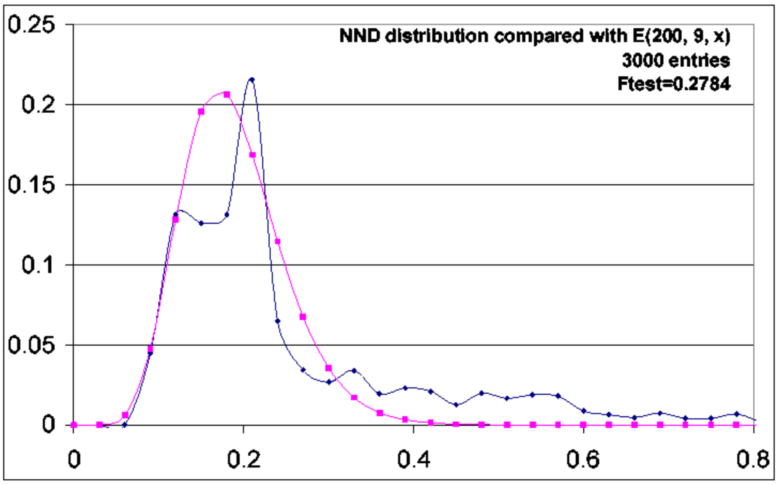

Fig. E200-9-x.ps. Comparison of Erlang distribution with NND^2 distribution in E10.

Nearest neighbour ball volume with respect to the index

Very interesting question in problem studied is dependence of the 1st, 2nd, 3rd,... nearest neighbour ball size (or the 1st, 2nd, 3rd,... nearest neighbour distance to the *n*-th power) wrt. the index of the corresponding nearest neighbour. Usually it would be supposed that this dependence would be strictly linear. It can be shown, that it is not true for any Euclidean space dimension $n > 1$.
For low dimensions the difference between the distances to the same power as is the space dimensionality n and linear dependence was computed. The distance to the same power as is

the space dimensionality n was used as it is proportional to the volume of a ball in that space. For *n* = 1 only error curve was obtained, no systematic dependence is seen.

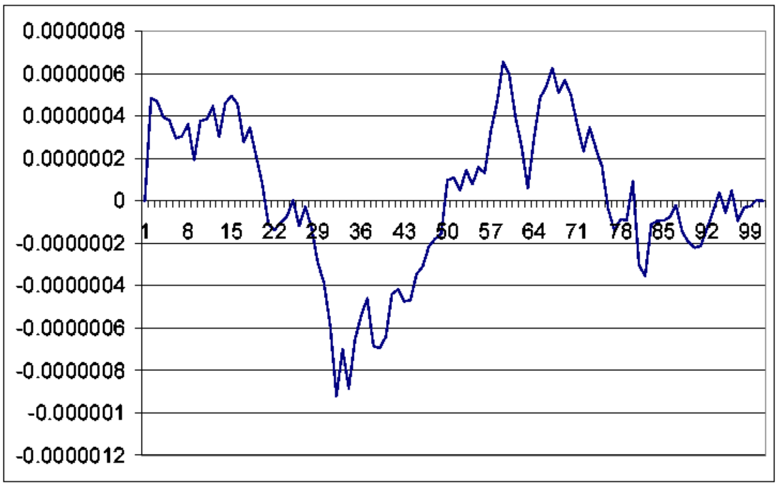

Fig. D^1E1avg-lin.ps. The difference between nearest neighbour No. 1...100 distances and linear dependence in E1. The curve shows, in fact, errors only.

For higher dimensions there is systematic difference which shows that volume of balls corresponding to the 1st, 2nd ... nearest neighbour in given Euclidean space does not grow linearly but with a little larger speed. It is best seen in Fig. D^10E10avgMed for E10.

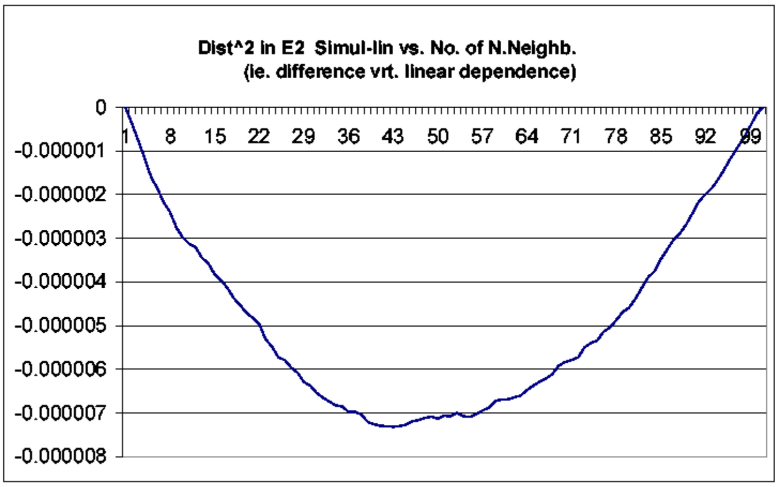

Fig. D^2E2avg-lin.ps. The difference between nearest neighbour No. 1...100 squared distances and linear dependence.

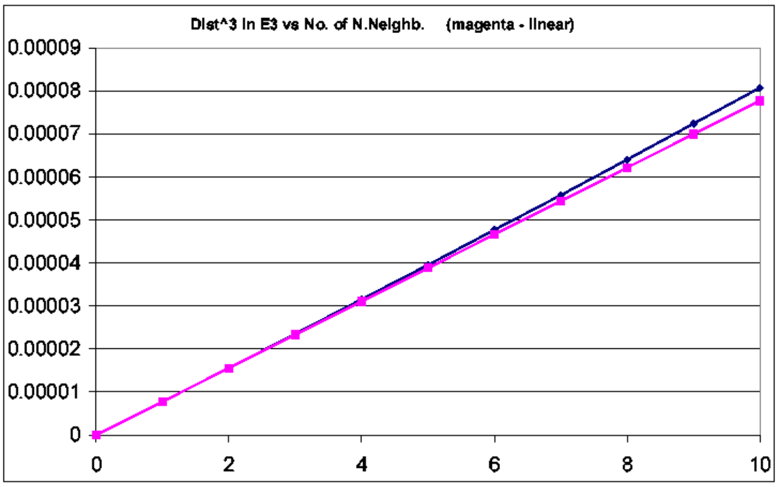

Fig. D^3E3avg-lin.ps. The nearest neighbour No. 1...10 distances to the third power and linear dependence.

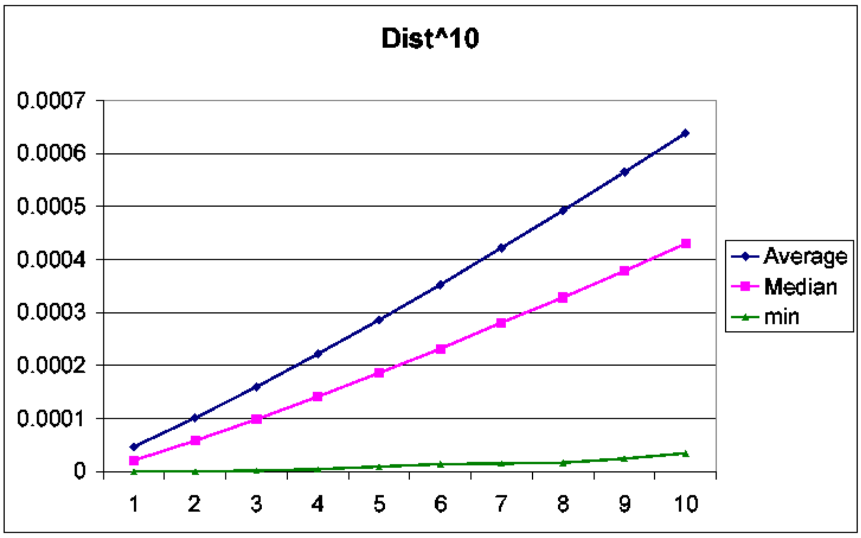

Fig. D^10E10avgMed.ps. The nearest neighbour No. 1...10 distances to the tenth power: average, median and minimum.

Sum of inverted volumes

Interesting thing is if the following sums are convergent or not

$$S_1 = \sum_{i=1}^{\infty} \frac{1}{d_i^n} \quad \text{and} \quad S_1 = \sum_{i=1}^{\infty} \frac{1}{d_i^{n-1}} \quad . \tag{1}$$

From the preceding part it is seen that d_i^n as well as d_i^{n-1} for $n > 1$ grow rapidly than proportionally to i, i.e. rapidly than denominators of divergent harmonic series $1/i$, which is considered as an threshold between convergent and divergent serii. Practical considerations show that the convergence is sufficient.

NND to n-th power increment

In the following figures are shown the increments of some powers of distances of first ten nearest neighbours in E10. Mostly the means are shown of 32000 samples. In each figure the first column shows the corresponding power of the nearest neighbour distance the others are differences.

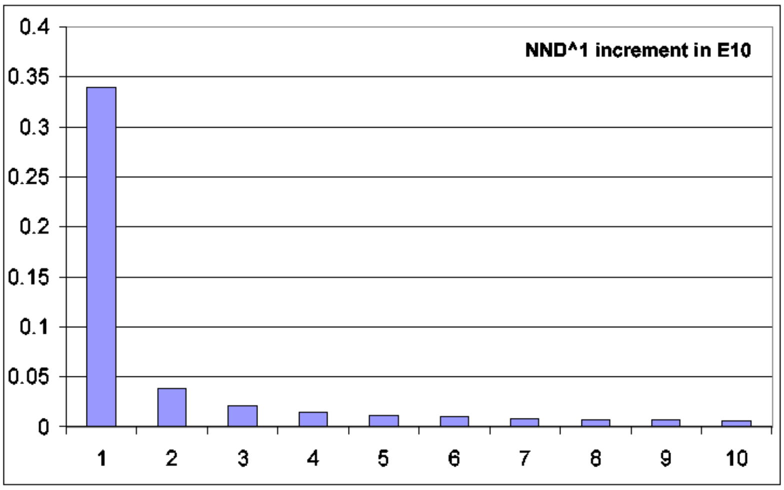

Fig. D^1incrE10.bmp. The nearest neighbour No. 1...10 distance differences in E10.

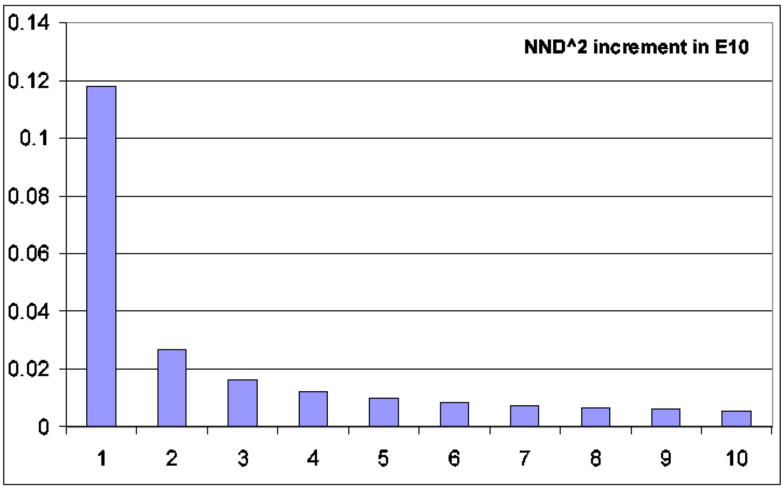

Fig. D^2incrE10.bmp. The nearest neighbour No. 1...10 distances squared differences in E10.

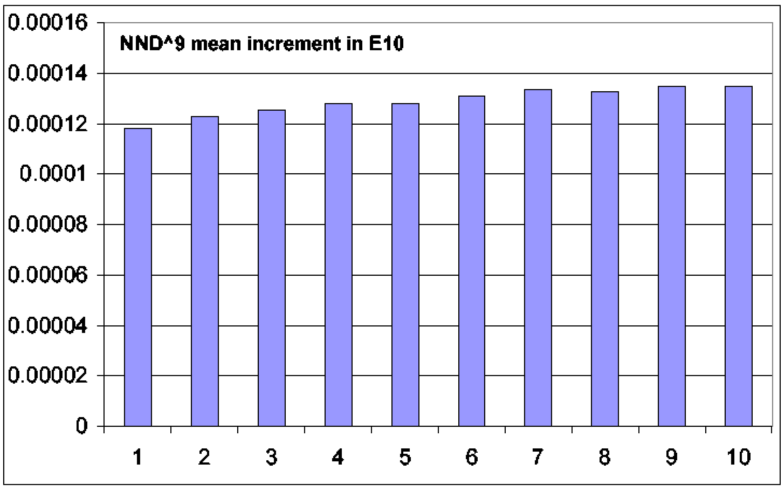

Fig. D^9incrE10mean.bmp. The nearest neighbour No. 1...10 distance to the 9th power differences in E10.

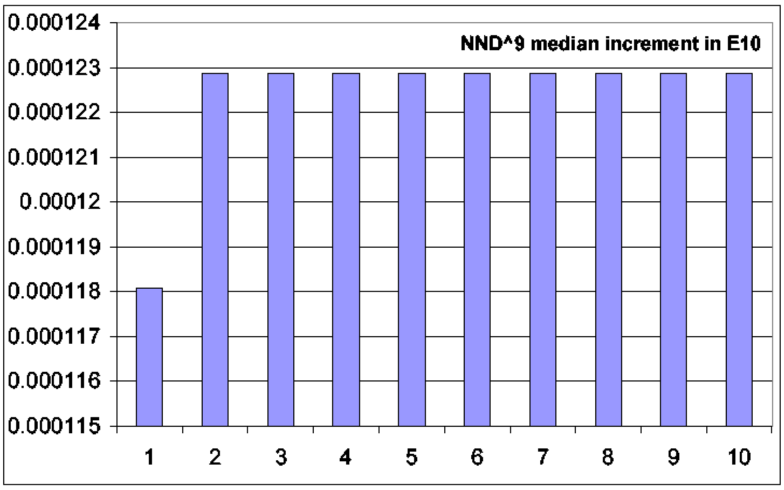

Fig. D^9incrE10medi.bmp. The nearest neighbour No. 1...10 distance to the 9th power differences in E10 - medians.

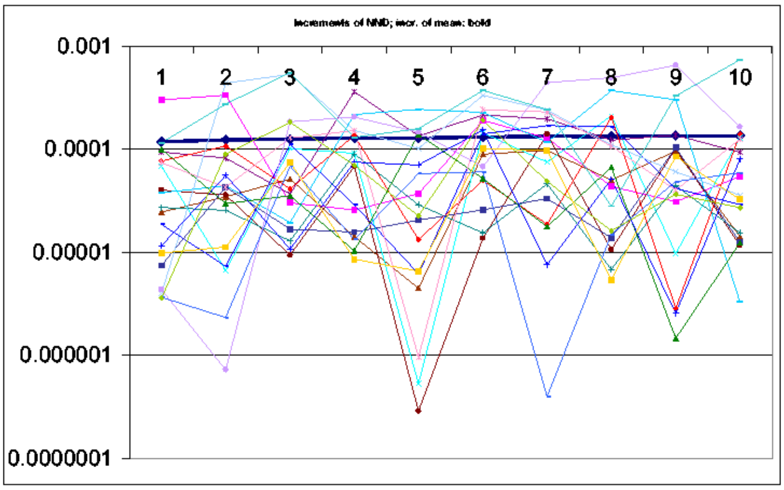

Fig. D^9incrE10xx.bmp. The nearest neighbour No. 1...10 distance to the 9th power differences in E10. The bold line shows mean, the others are individual randomly selected cases.

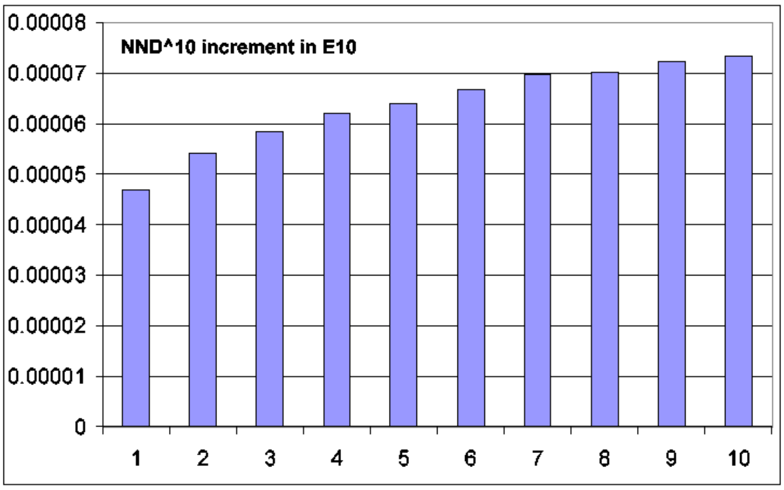

Fig. D^10incrE10.bmp. The nearest neighbour No. 1...10 distance to the 10th power increments in E10.

It is interesting that for (n-1)th power in *En* the differences are in average (mean) nearly constant. There is also the fact that the differences of the n-th power of 1st, 2nd, 3rd ... neighbours distance, i.e. volume of correspondig "betweenballs" grow monotonously. As each "betweenball" corresponds to just one point in the space *En* then the density of the space is from this point of view graduallu lessens, the space gets more and more thin with distance from any of our fixed point.

Conclusions

The target of this study was to get clear answer if the series (1) can be considered as convergent or not. It was shown that there is a convergence but it is not very fast in Euclidean space with randomly distributed points with homogenous distribution. Very positive fact is that the convergence does not depend on space dimensionality *n*.

Acknowledgement

This work was supported by the Ministry of Education of the Czech Republic under project No. LN00B096.

References

[1] www.xycoon.com/erlang.htm
[2] www.ee.aston.uk/teaching/tutorials/traffic/ErlangDistribution
[3] Lee, P.M.: Bayesian Statistics, an introduction. Oxford University Press New York, 1988.
[4] Refregier, P., Vallet, F.: Probabilistic Approach for Multiclass Classification with Neural Networks. In: Artificial Neural Networks – Proc. of the 1991 Int. Conf. on Artificial Neural Networks (ICANN-91), Espoo, Finnland 24-28 June, 1991, Ed. T. Kohonen et al., North-Holland Amsterodam, 1991, Vol 2, 1003-1006.