UDC 614.2:517.004.13:519.46(014)

O.M. Ocheredko, V.P. Klimenyk

National Pirogov Memorial Medical University, Vinnytsya

USING OF THE SEMIVARIOGRAM TO MODEL GEOGRAPHICAL DISTRIBUTION OF HEALTH EVENTS (EXAMPLE OF CARDIOVASCULAR DISEASES DISABILITY RATES IN ZHYTOMYR REGION OF UKRAINE)

Geostatistical modelling is on the rise of its popularity. Still limited tools are available to model health events. The potential of semivariogram to validation of geostatistical modelling was unveiled. Age-period-cohort (APC) data are 11 birth cohorts from date of birth before 1940 year and consequently by five year intervals, 13 443 498 adult-years totally. Data covers all (26) counties of Zhytomyr region, Ukraine. Incidence disability cases certified in 1999–2008 were retrieved from records of medical expert committees. It is used methods: semivariogram modelling with spatial covariance matrix processed by VARIOGRAM and GLIMMIX procedure, SAS. It is demonsrated, that validated initial parameters to space heterogeneity model enable verify robust spatial covariance matrix to study the irregularity in distribution of APC-factors.

Key words: *semivariogram, disability, cardiovascular diseases, geostatistical modelling, Ukraine.*

Geostatistical modelling of health events exploits continuous spatial framework that borrows strength from interpolation between observed point readings. The simplest approach is to integrate spatial correlations in covariance matrix [1]. Bayesian technics exploiting spatial priors is advanced approach [2]. Cardiovascular diseases proved to have distinct population distributions both cross population groups and by localities [3]. Still CD-related disability being the notorious leader somehow evaded attention of geostatistical researchers. Unfortunately we failed to find such researches for CD-related disability in particular in Ukraine.

Materials and Methods. The frame of study was shaped by spatial data organization by APC-factors. It renders «growth» of birth cohorts in time and age. The principal unit is birth cohort. We studied 11 cohorts from date of birth before 1940 and consequently by five year intervals («1941–1945», «1946–1950», … , «after 1985»), 13443498 adult-years totally (table 1). Each cohort captures unique combination of historical events [4]. The other important component is time dimension that unfolds the succession of events. Age has intrinsic importance per se as well as indispensable covariate to solve ambiguity of time-age collinearity. Counties have been described by sociodemographic variables and geographical coordinates. More to design can be found in [5–7].

Using semivariogram.

We enhanced classical approach to the geostatistical modelling. It suffers from inconsistencies of starting values of parameters. That is why we suggest the computation of sample or empirical measures of spatial continuity. These continuity measures are the regular semivariogram, a robust version of the semivariogram, and the space covariances. Spatial prediction, then, involves two steps. First, we model the covariance or semivariogram of the spatial process. This involves choosing both a mathematical form and the values of the associated parameters. Second, we use this dependence model in solving the kriging system at a specified set of spatial points, resulting in predicted values and associated standard errors.

We estimated the spatial variability and used the estimates to smooth observed spatial disabi-

Table 1. Composition of birth

Cohorts	Years				
	1999	2000	2001	2002	2003
Before 1940	400000	384000	317000	317000	–
1941	71422	78418	61580	60356	373000
1946	91464	92175	82294	84398	85637
1951	98822	99236	91768	92091	94274
1956	98951	95759	99698	99974	98434
1961	90980	92262	94107	91655	88479
1966	101000	100000	91425	91131	90661
1971	99644	99350	95693	94689	90661
1976	379000	372000	92123	92555	91432
1981	–	–	352000	346000	329000
After 1985	–	–	–	–	–
Totals	1431283	1413200	1377688	1369849	1342759

lity rates. Smoothing is especially propitious for small communities due to sporadic fluctuation of disability rates. Next, we used smoothed disability rates to examine plausible input to space heterogeneity of APC-factors.

The empirical semivariogram was computed by classical estimator from data residuals (r) using the formula:

$$\gamma(h) = \frac{1}{2m} \sum (r_i - r_j)^2 ,$$

where m is the number of pairs of observations a distance, h, apart, $\gamma(h)$ is estimated for all distances at which pairs of observations exist or at a discrete set of lagged values within a tolerance to ensure that a sufficient number of observations contribute to each value of $\gamma(h)$. The semivariogram was computed using the following SAS statements (LP stands for linear predictor):

```
proc glimmix data=spacedata;;
class county cohort year Gender Residence;
model Disabilitynum/population=LP;
output out=variogramdata resid (noblup ilink)=r_ilink;
id latitude longitude;
run;
proc variogram data = variogramdata outvar=outv;
coordinates x=longitude y=latitude;
compute novariogram;
var r_ilink;
run;
proc variogram data = variogramdata outvar=outv;
coordinates x=longitude y=latitude;
compute lagd=10 maxlag=14;
var r_ilink;
run;
```

Procedure GLIMMIX was used to get residuals r_ilink as input to procedure variogram. First run of latter supplied the value for basic distance unit defining the lags (LAGD=) and maximum number of lag classes used in constructing the continuity measures (MAXLAG=). The average length of the lag is around 10 and final filled so lagd=10 (fig. 1). Up to a pairwise distance of 140 we have a sufficient number of pairs. With choice of LAGD=10, this yielded a maximum number of lags =140/10=14. Fig. 2 shows that the choice of a semivariogram model is adequate. Two functions (exponential and Gaussian) regressed on values of sample semivariogram (variable «variog») with parameters range (R)=21 NUGGET=0.047 SILL=0.05 were fit by NLIN procedure:

```
proc nlin data= variogramdata;
parms R=50.5 NUGGET=0.047 SILL=0.052;
model variog = NUGGET*(distance=0) + SILL*(1-exp(-distance/(r)))*(distance<r) + SILL*(distance>=r);
run;
proc nlin data= variogramdata;
parms R=21 NUGGET=0.047 SILL=0.05;
model variog = NUGGET*(distance=0) +
```

cohorts by years

2004	2005	2006	2007	2008	Totals
–	–	–	–	–	1418000
304000	294000	308000	280000	–	1830776
66114	72768	75512	123420	152580	926362
85651	86041	86019	77085	78791	889778
92967	95177	94988	85909	87844	949701
95571	92043	89876	95002	93628	923603
87250	87572	87191	87351	84896	908477
91007	90765	90621	87615	87556	928782
90477	89089	87698	89346	88296	1472016
414000	95771	99338	87952	89942	1814003
–	310000	294000	394000	384000	1382000
1327037	1313226	1313243	1407680	1147533	13443498

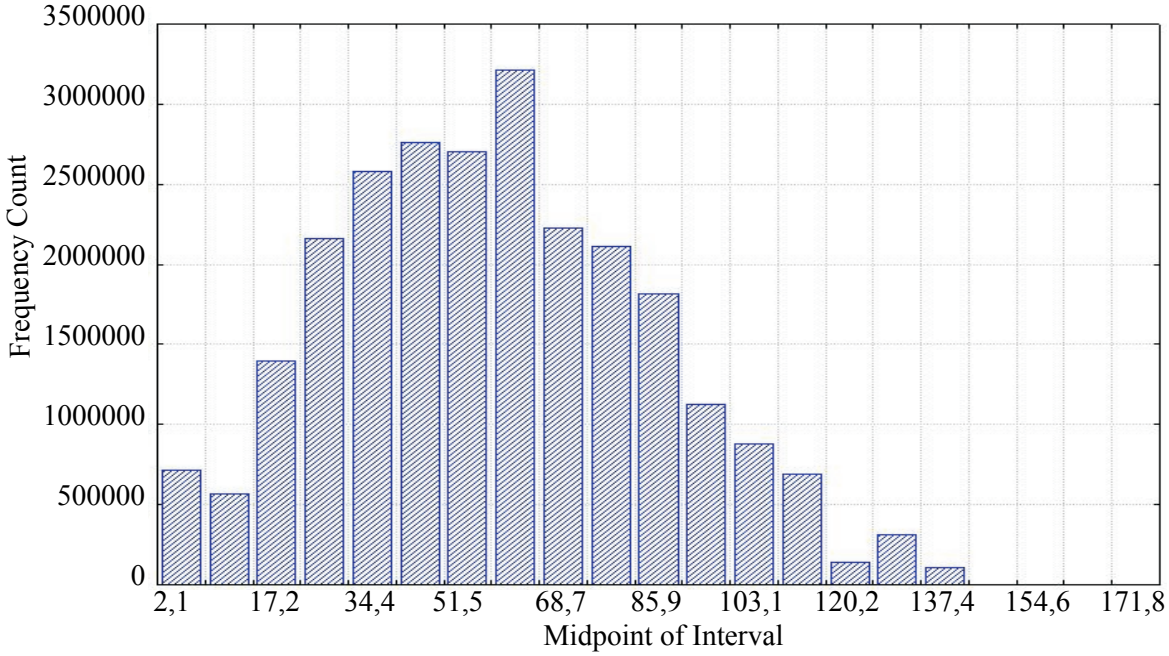

Fig. 1. Distribution of Pairwise Distances

SILL*(1–exp(–distance*distance/(R*R)))*(distance>=R)+ SILL*(distance>=R);
run;

We used Gaussian function with these particular parameters in KRIGE2D procedure to produce a contour plot of the kriging estimates and the associated standard errors (r~2*range=100′=1°40′ for Gaussian covariance structure):

proc krige2d data= variogramdata outest=est;
pred var= r_ilink k r=100;
model NUGGET=0.047 SCALE=0.052 RANGE=50.5 form=gauss;
coord xc= latitude yc= longitude;
grid x=2960 to 3100 by 5 y=1640 to 1800 by 5;
run;

Grid values ranges from 49°20′ to 51°40′ for latitude (given in minutes) and from 27°20′ to 30°00′ for longitude. NUGGET, SCALE, RANGE defined by variogram (fig. 2).

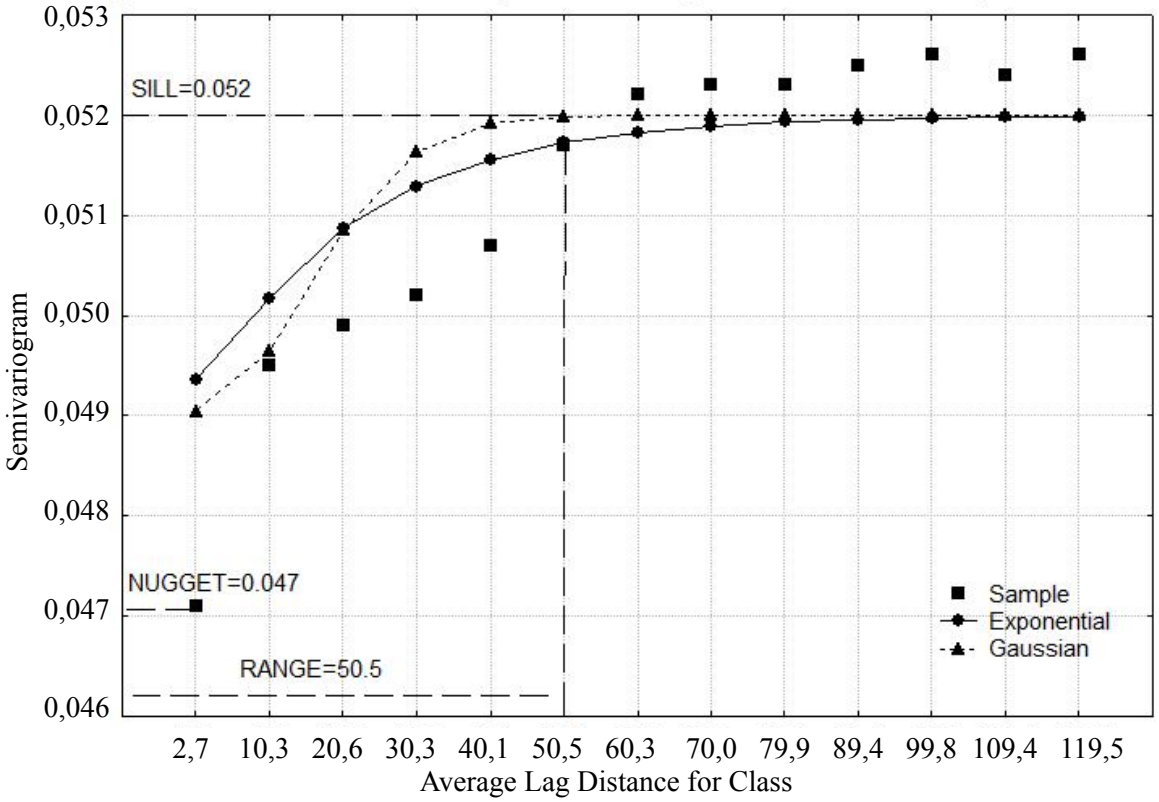

Fig. 2. Theoretical and Sample Semivariograms for CD Disability Data

Precise estimations of space covariance matrix random effects NUGGET, SILL, RANGE we received from procedure MIXED. We recommend to use it with the scope of initial values around variogram estimates via PARMS option. Otherwise, procedure can converge to local maxima and provide grossly unreasonable estimates. Sequence of initials should correspond to the given in output list:

```
proc mixed data= variogramdata;
class county;
model r_ilink =;
repeated/ subject = county
type = SP(GAU)(latitude longitude)
local;
parms     (0 to 0.5 by 0.02)
          (0 to 100 by 5)
          (0.045 to 0.049);
run;
```

The random effects estimates were: SILL=0,050, RANGE=20,7 NUGGET=0,047. These defined space covariance matrix.

Modifications to space heterogeneity in disability rates by irregularity in distribution of APC-factors were assessed by comparison of likelihoods of nested models. Models differed by linear predictor (LP): shell model with only intercept in LP while LP of APC model was padded out with APC factors. Models had the same spatial Gauss form covariance matrix. Binominal models with canonical logit link function were processed by procedure GLIMMIX.

```
proc glimmix data=spacedata;
class county cohort year Gender Residence;
model Disabilitynum/population=LP;
random _residual_ / subject = intercept
type = SP(GAU)(latitude longitude);
parms (0.050 20.7 0.047)
run;
```

Results. The original data in Zhytomyr region bear evidences of heterogeneity in space distribution of CD-related disability risks (fig. 3). Three picks are conspicuous. Listed by height they are: northern (of latitude above 51° within longitude of the range 27°20′–29°), southern (of latitude up to 50°40′ within longitude of the range 27°40′–29°40′), and eastern (in ranges of latitude 50°20′ – 51°00′ and longitude from 29°30′).

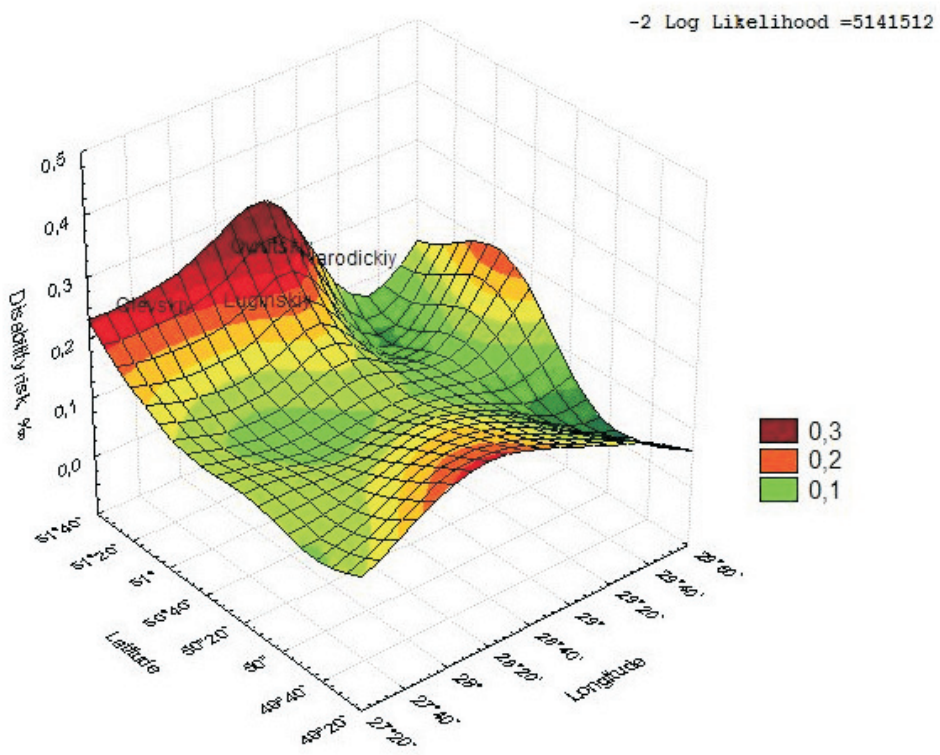

Fig. 3. Space distribution of CD-related disability risks (Zhytomyr region, Ukraine)

Possible applications.

Testing of hypothesis that space heterogeneity in disability rates enhanced by irregularity in distribution of APC or other potent factors can be proceeded by unwrapping the corresponding trend fixed effect. Say, if we are interested in test of age input to heterogeneity we compare to models with the same covariance matrix but different LP (with age + squared age terms and with these dropped out). If space distribution alters significantly, we infer that age indeed modifies space heterogeneity. We have used the difference between models likelihoods as operational test statistics. The procedure applied to test consequently the input of other factors. Basically, we compare 2 consequent nested models: shell model and hypothesized factor model.

Shell model. It is composed by space covariance structure (zero ground model). The «subject = intercept» option in procedure GLIMMIX treats all observations in the data set as potentially correlated. In fact, shell model depicts somewhat smoothed original space distribution of CD-related risks of disability. Negative double logarithm of shell model likelihood (-2 Log Likelihood) equals 5141512.

APC -(factor) model. Including APC-related components to LP covariance matrix preserved helps to test the significance of induced changes to space distribution of risks. LP now contains APC variables cohort, year, age, age squared. -2 Log Likelihood now dropped to 3518939.

Number of additional fixed parameters against shell model is 21. The difference between values of double logarithms of model likelihoods follows in approximation chi-square distribution with degrees of freedom equals to number of additional parameters. Δ(2 Log Likelihood) = (2 Log Likelihood$_{shell}$ – 2 Log Likelihood$_{APC}$=5141512 – 3518939 =1622573). Chi-square(21) 0,999 centile = 54 that is considerably less then difference Δ(2 Log Likelihood). Therefore, we stipulate significant $p<0,0001$ impact of APC-factors on geographical distribution of CD-related risks of disability in Zhytomyr region, Ukraine.

Conclusions

1. Proposed 2-step approach based on semivariogram proved to be indispencible to geographical modelling of health events.

2. Geographical distribution of CD-related risks of disability for Ukraine is still to be

studied as well as the problem of disability on a whole.

3. Hindrances posed by complex data structure (Fu W. J., 2000). The other relates to inconsistency of registries that calls for smoothing.

4. Geostatistical modelling may be helpful in adjustment of local minor irregularities by smoothing as well as in explanation gross discrepancies.

5. Behind seemingly stationary time distribution of CD-related disability risks we discovered significant trends and distributions by cohorts, years, age by doing APC-decomposition in space frame.

References

1. SAS for mixed models / [Littell Ramon C., George A. Milliken, Walter W. Stroup et al.]. – [2nd ed.]. – Cary, NC : SAS Institute Inc., 2006. – 828 p.

2. *Congdon P. D.* Applied Bayesian Hierarchical Methods / P. D. Congdon. – Chapman and Hall/CRC, 2010. – 604 p.

3. Individual social class, area based deprivation, cardiovascular disease risk factors, and mortality: the Renfrew and Paisley Study / G. D. Smith, C. Hart, G. Watt [et al.] // J. Epidemiol. Community Health. – 1998. – V. 52. – P. 399–405.

4. *Fu W. J.* Ridge estimator in singular design with application to age-period-cohort analysis of disease rates / W. J. Fu // Communications in Statistics-Theory and Method. – 2000. – V. 29 (2). – P. 263–278.

5. *Klimenyk V.* The study of cardiovascular disability risks by the birth cohorts / V. Klimenyk // East European J. of Pub. Health. – 2013. – V. 1 (21). – P. 155–156.

6. Analysis of CD related disability risks by APC construction on the basis of strip-split-plot design / V. Klimenyk, O. Galachenko, O. Ocheredko, I. Andrievsky // East European J. of Pub. Health. – 2012. – V. 2–3 (18–19). – P. 155–163.

7. *Klimenyk V.* The investigation of historical trends of stroke disability rates based on APC-decomposition / V. Klimenyk // East European J. of Pub. Health. – 2013. – V. 1 (21). – P. 63–71.

О.М. Очередько, В.П. Клименюк

ВИКОРИСТАННЯ СЕМІВАРІОГРАМИ ДЛЯ МОДЕЛЮВАННЯ ГЕОГРАФІЧНОГО РОЗПОДІЛУ МЕДИЧНИХ ПОДІЙ (НА ПРИКЛАДІ РІВНІВ ІНВАЛІДНОСТІ ВНАСЛІДОК СЕРЦЕВО-СУДИННИХ ЗАХВОРЮВАНЬ У ЖИТОМИРСЬКІЙ ОБЛАСТІ УКРАЇНИ)

Геостатистичне моделювання набуває все більшої популярності. Проте цей процес стримує відчутна нестача програмного забезпечення, зокрема для моделювання медичних подій. Розкрито потенціал семіваріограми для валідизації геостатистичних моделей медичних подій. Організовано за APC-дизайном (структуровані за віком, періодом, когортою за народженням) 11 когорт за датою народження до 1940 року і далі за 5-річними інтервалами, всього 13 443 498 чоловіко-років. Дані щодо випадків інвалідизації отримано суцільним методом по 26 районах Житомирської області на основі актів освідчення на ВТЕК/МСЕК. Використовували методи: побудови варіограми та геостатистичного моделювання за просторовою коваріційною матрицею на основі процедур VARIOGRAM і GLIMMIX, SAS. Показано, що валідизовані на основі варіограми параметри просторової коваріційної матриці уможливлюють вивчення просторової гетерогенності розподілів, зокрема APC-факторів. Геостатистична модель дозволила розкрити суттєві тренди і розподіли ризиків інвалідизації в розрізі когорт, років, вікових груп населення на основі APC-декомпозиції в просторовому фреймі.

Ключові слова: *інвалідизація, серцево-судинні захворювання, семіваріограма, географічна модель.*

А.Н. Очередько, В.П. Клименюк

ИСПОЛЬЗОВАНИЕ СЕМИВАРИОГРАММЫ ДЛЯ МОДЕЛИРОВАНИЯ ГЕОГРАФИЧЕСКОГО РАСПРЕДЕЛЕНИЯ МЕДИЦИНСКИХ СОБЫТИЙ (НА ПРИМЕРЕ УРОВНЕЙ ИНВАЛИДНОСТИ ВСЛЕДСТВИЕ СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ В ЖИТОМИРСКОЙ ОБЛАСТИ УКРАИНЫ)

Геостатистическое моделирование приобретает все большую популярность. Однако этот процесс сдерживает ощутимый дефицит программного обеспечения, в частности для моделирования

медицинских событий. Раскрыт потенциал семивариограммы для валидизации геостатистических моделей медицинских событий. Организованы по APC-дизайну (структурированы по возрасту, периоду, когорте по рождению) 11 когорт по дате рождения до 1940 года и позже по 5-летним интервалам, всего 13 443 498 человеко-лет. Данные относительно случаев инвалидизации получены сплошным методом по 26 районам Житомирской области на основе актов освидетельствования на ВТЭК. Использованы следующие методы: построения вариограммы и геостатистического моделирования на основе пространственной ковариационной матрицы с помощью процедур VARIOGRAM і GLIMMIX, SAS. Показано, что валидизированные вариограммой параметры пространственной ковариационной матрицы позволяют исследовать пространственную гетерогенность распределений, в частности APC-факторов. Геостатистическая модель позволила раскрыть существенные тренды и распределения рисков инвалидизации в разрезе когорт, лет, возрастных групп населения на базе APC-декомпозиции в пространственном фрейме.

Ключевые слова: *инвалидизация, сердечно-сосудистые заболевания, семивариограмма, географическая модель.*