УДК 519.6

Исследование обобщённых смешанно-аддитивных регрессионных моделей с пространственно-структурными факторами рисков

Е. Ю. Щетинин

Кафедра прикладной математики
ФБГОУ ВПО МГТУ «СТАНКИН»
Вадковский пер., д. 3а, Москва, Россия, 119136

В настоящей работе реализован байесовский подход к решению задачи распространения эпидемии опасных инфекционных заболеваний на основе обобщённых смешанно-аддитивных регрессионных моделей с пространственно-временными факторами. В модель одновременно включены как непрерывные, так и категориальные переменные, а также структурные эффекты, учитывающие влияние статистических связей на риски инфицирования. В качестве основных таких переменных были обоснованно выбраны наличие густонаселённых районов малоимущих с низким уровнем медицинского и социального обеспечения, плотность населения в них. В качестве непрерывных факторов выбраны близость проживания к свалкам, плотность расположения свалок, а также близость проживания к потенциальным источникам заражения холерой. Для оценивания параметров модели нами использованы байесовские сплайны, а также марковские случайные поля как стохастический аналог многомерных пространственных структур связей регрессоров. На примере эпидемиологических данных заболевания холерой в Гане нами проведены вычислительные эксперименты по оцениванию различных характеристик поражения населения холерой, дан прогноз по распространению заболевания по территории страны и численности заражённых. Сравнительный анализ предложенной модели и классических регрессионных моделей показал её вычислительную эффективность и высокую точность в оценках прогноза риска инфицирования.

Ключевые слова: эпидемиологический риск, холера, байесовский полупараметрический подход, регрессионная модель, категориальный регрессор, марковские цепи, пространственно-аддитивные структуры, марковские случайные поля, сплайны, прогноз.

1. Введение

Проблема разработки моделей, описывающих, описывающих риск, опасность наступления некоторых неблагоприятных событий, таких как массовое заражение опасными вирусными инфекциями (эпидемии), смертельные исходы в результате перенесённых тяжёлых заболеваний (онкология инсульт, инфаркт и пр.), экология (загрязнение окружающей среды опасными отходами жизнедеятельности человека, непредсказуемые природные катаклизмы), экономика (прогнозирование дефолтов), учитывающие разноплановые факторы, оказывающие на них ключевое влияние, а также их взаимодействие в пространственно-временном диапазоне, является современной и актуальной. Приложения таких моделей и методов их анализа всегда востребованы в самых разнообразных областях жизнедеятельности человека.

История развития подобных моделей начинает свой отсчёт в работах [1, 2] и связана с обобщением аддитивных и смешанных регрессионных моделей. Структурные аддитивные модели стали следующим этапом их развития, позволяющие включить в рассмотрение пространственные эффекты на малых площадях. В настоящей работе нами развивается и расширяется этот подход на факторы различного типа, такие как категориальные и дискретные, в едином формате методов анализа модели [3]. С целью включения в модель как нелинейных и непрерывных эффектов, так и пространственно-временных факторов, нами использован подход обобщённых полупараметрических смешанных регрессионных моделей [4]. Пусть

Статья поступила в редакцию 13 апреля 2014 г.

известны наблюдения $(y_i, \mathbf{z}_i)$, $i = 1, ..., n$, где y_i — отклик на вектор $\mathbf{z}_i$, содержащий факторы различного типа. При этом будем полагать, что условное математическое ожидание имеет вид $M(Y|\,\mathbf{z}) = \psi(\eta(\mathbf{z}))$, где $\psi(\cdot)$ — некоторая функция, свойства которой будут сформулированы позже. Под риском, опасностью наступления некоторых неблагоприятных событий будем понимать случайную величину $Y = \{y_i\}, i = 1, ..., n$ со следующими свойствами: $y_i \approx F(\eta_i, \sigma^2)$, F_i — непрерывная функция распределения y_i неизвестное среднее η_i выражается как

$$\eta_i = \sum_{k=1}^{p} f_k(x_{i,k}) + f_{spat}(s_i) + \omega_i\gamma, \tag{1}$$

где $f_k(x_{i,k})$ — функции непрерывных факторов $x_{i,k}$ и $f_{spat}(s_i)$ — функция пространственных связей между различными географическими регионами s_i, $i = 1...S$ области, на которой исследуется модель (1). Далее, эту функцию разбивают на две компоненты, учитывающие как структурные эффекты связей различных факторов на области S так и пространственно некоррелированные эффекты

$$f_{spat}(s) = f_{str}(s) + f_u(s).$$

Третий компонент модели (1) представляет собой вектор $\omega_i = (\omega_{i,1}, ..., \omega_{1,r})$ категориальных факторов, где γ — параметр. При исследовании свойств отклика η_i будем полагать, что его распределение F может быть гауссовым, биномиальным (полиномиальным) или пуассоновским.

2. Метод байесовских сплайнов

Для моделирования функций f_k воспользуемся методом байесовских P-сплайнов [5, 6]. Этот метод предполагает, что неизвестные функции $f_k(x)$ могут быть аппроксимированы полиномиальным сплайном степени l, заданным на сетке $x_j^{\min} = \xi_{j,0} < \xi_{j,1} < ... < \xi_{j,s} = x_j^{\max}$. Такие сплайны могут быть записаны в терминах линейной комбинации базисных функций B_m

$$f_j(x_j) = \sum_{m=1}^{d} \varsigma_{j,m} B_m(x_j), \quad d = s + l. \tag{2}$$

Оценивание функций f_j сводится таким образом к вычислению вектора регрессионных параметров $\varsigma_j = (\varsigma_{j,1}, ..., \varsigma_{j,m})$, а выбор количества узлов производится в соответствии с методом, предложенным в работе [7], чтобы в достаточной степени обеспечить точность приближения данных и, вместе с тем устойчивость решения. Обычно в байесовском подходе штрафные сплайны вводятся путём замены производных штрафных функций их стохастическими аналогами, а именно случайным блужданием первого или второго порядка для регрессионных коэффициентов. В частности, случайное блуждание 1-го порядка имеет вид

$$\varsigma_{j,m} = \varsigma_{j,m-1} + u_{j,m}, \quad m = 2, ..., d, \tag{3}$$

а случайное блуждание 2-го порядка

$$\varsigma_{j,m} = 2\varsigma_{j,m-1} - \varsigma_{j,m-2}, \quad m = 3, ..., d,$$

где $u_{j,m} \approx N(0, \tau_j^2)$ — белый шум.

Условные распределения регрессионных параметров $\varsigma_{j,m}$ для случайного блуждания 1-го порядка имеет вид

$$(\varsigma_{j,m} \,|\varsigma_{j,m-1}) \approx N\left(\varsigma_{j,m-1}, \tau_j^2\right) \tag{4}$$

и для случайного блуждания 2-го порядка

$$(\varsigma_{j,m} \,|\varsigma_{j,m-1}\,, \varsigma_{j,m-2}) \approx N\left(2\varsigma_{j,m-1} - \varsigma_{j,m-2}, \tau_j^2\right). \tag{5}$$

Совместное распределение вектора $\varsigma_j = (\varsigma_{j,1}, ..., \varsigma_{j,m})$ предстаёт в виде многомерного гауссового распределения $p\left(\varsigma \left|\tau_j^2\right.\right) \approx \exp\left(-\frac{\varsigma_j' K_j \varsigma_j}{2\tau_j^2}\right)$, где ковариационная матрица K_j фактически выступает в роли штрафной функции. Равновесное соотношение точности и гладкости решения обеспечивается параметром вариации τ_j^2: чем больше его значение, тем сильнее сглажено решение, и наоборот.

3. Моделирование пространственно-временных факторов

Рассмотрим методы моделирования и оценивания пространственных факторов модели (1). В работе нами используется модель ближайших соседей на основе марковских случайных полей [8,9]. Итак, пусть вектор $s \in \{1, .., s_j, ., S\}$ содержит положения областей проживания или расположения объектов исследования на географическом полигоне S. Тогда локальную пространственную структуру статистических связей $f_{str}(s)$ между s_j и s_j' можно определить следующим образом

$$f_{str}(s) \,|f_{str}(s')\,,\ s \neq s',\ \tau_{str}^2 \approx N\left(\frac{1}{N_s}\sum_{s' \in \partial s} f_{str}(s'), \frac{\tau_{str}^2}{N_s}\right), \tag{6}$$

где N_s — число прилегающих к другим областям s, а $s' \in \partial s$ означает, что область s' является смежной с областью s Условное среднее функции $f_{str}(s)$ есть средневзвешенное оценок аналогичных функций смежных областей. Полагая, что эффект пространственных связей зависит напрямую от расстояния между центрами областей, для его описания мы выбрали модель k ближайших соседей. Это отражено в матрице смежности, подобно непрерывным функциям $f_j(x)$, баланс между точностью и гладкостью аппроксимации определяется параметром τ_{str}^2.

Для неструктурированных пространственных эффектов мы предполагаем, что функции $f_{unstr}(s)$ являются гауссовыми случайными независимыми величинами

$$f_{unstr}(s) \left|\tau_{unstr}^2 \right. \approx N\left(0, \tau_{unstr}^2\right).$$

Распределения гиперпараметров $\tau_j^2,\ j = 1, ..., p, str, unstr$ остаются неизвестными, но можно предположить, что оно является обратным гамма-распределением $IG(a, b)$ с плотностью

$$p\left(\tau_j^2\right) \propto \left(\tau_j^2\right)^{-a_j - 1} \exp\left(-\frac{b_j}{\tau_j^2}\right), \tag{7}$$

$$a_j = b_j = 0,001.$$

Байесовский подход к оцениванию модели 3 основан на анализе апостериорного условного распределения $p(\theta \,|Y)$, где θ — вектор параметров. Для модели (1)–(7) оно имеет следующий вид

$$p(\theta|Y) \propto \prod_{i=1}^{n} L(y_i, \eta_i) \times \\ \times \prod_{j=1}^{p} \left[p\left(\varsigma_j \left|\tau_j^2\right.\right)\right] \times p\left(f_{str} \left|\tau_{str}^2\right.\right) p\left(f_{unstr} \left|\tau_{unstr}^2\right.\right) \prod_{j=1}^{r} p(\gamma_j)\, p\left(\sigma^2\right), \quad (8)$$

где $L(\cdot)$ — функция правдоподобия. Совместное распределение для параметров γ, вектора $(\varsigma_1, ..., \varsigma_p)$, $f_{str}(s)$, $f_{unstr(s)}$ предполагается многомерным гауссовым.

4. Исследование эпидемии холеры в Гане с использованием пространственно-структурной регрессии

В качестве практического примера применения построенной модели рассмотрим данные представленные Всемирной организацией здравоохранения (WHO) и международной географической системой GIS Статистическая выборка содержит 26924 случаев заболевания холерой за период с 1999 по 2005 г., из которых зафиксировано 620 случаев смерти. В качестве непрерывных факторов выбраны близость проживания к ***свалкам*** ρ_1, плотность расположения ***свалок*** ρ_2 а также близость проживания к потенциальным источникам заражения холерой ρ_3. Дополнительно в модель введены следующие категориальные переменные: χ_1 — наличие густонаселённых районов малоимущих с низким уровнем медицинского и социального обеспечения в s_i регионе: если таковые в нем имеются, то $\chi_1 = 1$, в противном случае $\chi_1 = 0$; а также плотность населения χ_2: если в регионе плотность населения средний по области, то $\chi_2 = 1$, если ниже его, то $\chi_2 = 0$. Введённые бинарные переменные были рекомендованы к использованию в работе [9] как доступные, так и очевидные и наиболее причастные к развитию эпидемии. Для каждого населённого пункта i, $i = 1, ..., N$ с численностью населения P_i зарегистрированное количество заболевших Υ_i будем полагать случайной величиной, имеющей распределение Пуассона с интенсивностью $\Psi_i = \mathrm{E}_i \cdot R_i$, где E_i — средняя заболеваемость по i-му населённому пункту, R_i — риск заболевания. Общепринято на практике полагать [10,11], что $E_i = R \cdot P_i$, где R — общий риск заболевания, полученный по всей выборке наблюдений как средневзвешенное

$$R = \sum_{i=1}^{N} \frac{\Upsilon_i}{P_i} \times \frac{P_i}{\sum\limits_{i=1}^{N} P_i}. \quad (9)$$

Выражение (9) можно интерпретировать как относительный риск заболевания по всей выборке наблюдений для всего региона исследований.

5. Вычислительные эксперименты и выбор оптимальной модели

В работе приведён сравнительный анализ вычислительной и описательной эффективности различных регрессионных моделей на основе информационных критериев *AIC, BIC*, что, как правило, позволяет получить достаточно полную картину адекватности анализируемых моделей. **Модель 1** представляет собой линейную множественную регрессию:

$$\eta_i = \beta_1\rho_1 + \beta_2\rho_2 + \beta_3\rho_3 + \gamma_1\chi_1 + \gamma_2\chi_2. \quad (10)$$

Модель 2 является обобщённо-аддитивной, что предполагает нелинейность влияния непрерывных факторов на отклик:

$$\eta_i = f_1(\rho_1) + f_2(\rho_2) + f_3(\rho_3) + \gamma_1\chi_1 + \gamma_2\chi_2. \tag{11}$$

Наконец, **модель** (1)–(7):

$$\eta_i = f_1(\rho_1) + f_2(\rho_2) + f_3(\rho_3) + f_{str}(s) + f_{unstr}(s) + \gamma_1\chi_1 + \gamma_2\chi_2. \tag{12}$$

Результаты численных расчётов параметров модели (1) показали, что в рамках методологии линейной регрессии влияние непрерывных факторов ρ_1, ρ_3, учитывающих пространственные эффекты, на апостериорное среднее отклика η_i оказалось чрезвычайно малым. Основной вклад в **модель 1** принадлежит бинарным факторам наличия трущоб χ_1 и высокой плотности проживающего в них населения χ_2, и вместе с тем ею некорректно оценено влияние пространственно-структурных факторов. Оценки влияния пространственно-структурных факторов, полученные для **моделей 2, 3**, приведены на рис. 1, 2. Из них следует, что, во-первых, они имеют нелинейный характер, во-вторых, риск заражения холерой резко возрастает с увеличением параметра плотности свалок ρ_2 на исследуемой территории, начиная со значений $\rho_2 \geqslant 1,5 - 1,6$. Кроме того, риск заражения снижается некоторым нелинейным образом при увеличении расстояния до источников заражения ρ_1, что является вполне очевидным фактом.

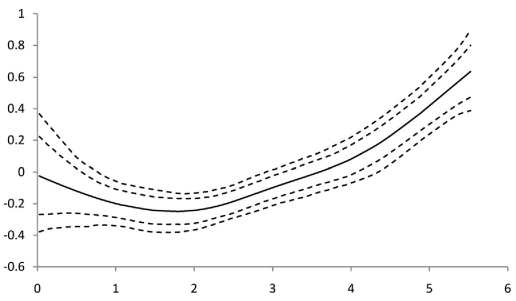

Рис. 1. График зависимости риска заражения холерой от плотности свалок $\eta(\rho_2)$

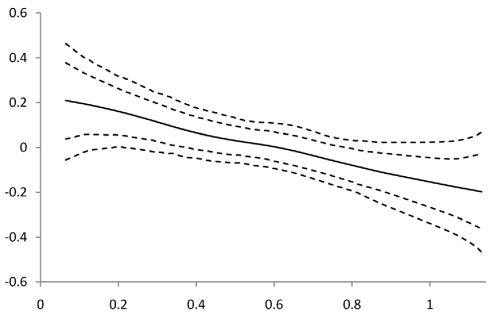

Рис. 2. График зависимости риска заражения холерой от расстояния до свалок $\eta(\rho_1)$

Для описания зависимости риска заражения холерой от расстояния до источника $\eta(\rho_1)$ оказалось достаточным использование линейного приближения (рис. 2).

На рис. 3 построены результаты оценивания структурных пространственных эффектов распределения риска заболевания холерой по географическим регионам. Области, окрашенные в тёмный цвет, соответствуют высокому уровню рисков заражения и высоким положительным значениям апостериорного среднего η. Области белого цвета соответствуют низкому уровню рисков заражения и отрицательным значениям η, серые указывают на области значений η, близких к нулю, т.е. на незначительные пространственные эффекты. Полученные результаты свидетельствуют о кластерности эпидемии холеры, с высоким уровнем риска заражения, наступающей в центральной части, и меньшего риска в юго-восточной части региона. Неструктурированные пространственные эффекты доминируют над структурированными пространственными эффектами. Об этом свидетельствует высокий коэффициент дисперсии.

Результаты проведённых исследований показали, что среди обитателей трущоб риск возникновения холеры особенно велик. Риск заражения также относительно высок в плотно населённых общинах. Хотя холера передаётся, главным

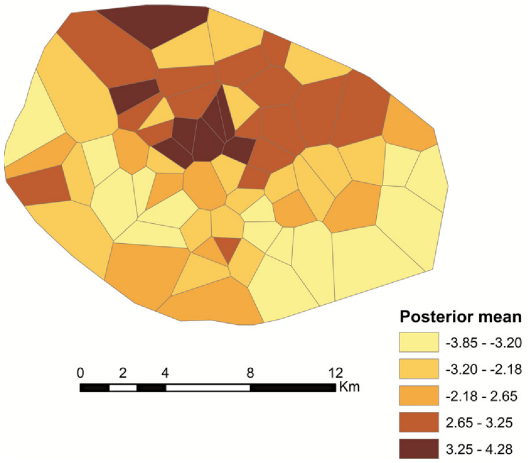

Рис. 3. Пространственное распределение апостериорных средних значений η по областям S

образом, через заражённую воду или пищу, неудовлетворительные санитарно-гигиенические условия, в среде обитания, могут повысить риски её распространения. Вирус холерных вибрионов может выживать и размножаться вне человеческого тела, а также быстро распространяться там, где условия жизни неудовлетворительны и отсутствует безопасная утилизация отходов жизнедеятельности человека. Кроме того, жители беднейших районов сталкиваются с первоочередными бытовыми проблемами, в том числе доступ к питьевой воде и средствам санитарии. Во многих случаях коммунальные службы не могут обслуживать этот контингент из-за факторов, связанных с юридическими аспектами системы землепользования, техническими правилами обслуживания. В большинстве районов трущобы также находятся в низко лежащих территориях, подверженных наводнениям. Особенности рельефа, почвенных и гидрогеологических условий делают трудно поддерживаемыми высокие стандарты санитарии среди таких жителей. Риск заражения холерой снижается с увеличением расстояния от свалки, тогда как жители менее 500 м от свалок наиболее уязвимы, что согласуется с выводами аналогичных исследований [10, 11]. Наблюдаемое снижение рисков заражения холеры с увеличением расстояния от загрязнённых поверхностей водных объектов, и значимая линейная зависимость между параметрами ρ_1 и ρ_3 (результаты регрессионного анализа $\beta = 0,582,\ R^2 = 0,67p < 0,001$) подтверждают эту гипотезу.

Холера развивается в основном за счёт негативных экологических и социально-экономических факторов. Полученные результаты указывают на то, что географически близкие поселения, как правило, имеют схожие риски. Структурированные пространственные эффекты, входящие в состав модели по сути являются суррогатной мерой ненаблюдаемой пространственной корреляции факторов риска холеры. Эти структуры чётко указывают на возможные ненаблюдаемые факторы риска холеры, которые могут быть глобальными или локальными. Результаты также дают чёткую картину высокой кластерности холеры, с высоким риском в центральной части и низким риском в юго-восточной части региона. Так, например, повышенный риск в центральной части может быть результатом влияния высокого ежедневного притока торговцев и гражданских специалистов из других общин центрального района. Доминирование неструктурированных пространственных эффектов над структурированными пространственными эффектами указывает, что ненаблюдаемые факторы больше являются локальными, чем глобальными. Например, такие пространственные вариации может вызывать вариативность социально-экономических характеристик домохозяйств. Таким образом, это даёт повод для дальнейших эпидемиологических исследований с

использованием дополнительной информации о домохозяйствах в пространственном масштабе в пределах области исследования.

6. Заключение

Возросший интерес к использованию пространственно-аддитивных структур в статистической эпидемиологии связан с насущной необходимостью определения факторов, повышающих риск заражения. В работе предложен байесовский полупараметрический подход к моделированию развития эпидемии опасных вирусных заболеваний на примере холеры в географическом регионе Кумаси, Гана. Это позволило провести совместный анализ нелинейных эффектов непрерывных регрессоров, пространственно структурированных эффектов, неструктурированных неоднородностей и фиксированных эффектов регрессоров. Предложенная в работе модель показала, что риск заражения холерой связан с условиями проживания в трущобах с высокой плотностью населения в них, близостью и плотностью размещения к свалкам, близостью к потенциально загрязнённым рекам и ручьям, а также с возможными ненаблюдаемыми факторами риска. Полученные количественные оценки прогноза распространения эпидемии холеры в регионе предоставляют его государственным и муниципальным органам здравоохранения возможность предотвращения дальнейшего распространения заболевания и предупреждения новых случаев заражения. Обнаружение и исследование ненаблюдаемых факторов риска а также количественная оценка их влияния является дальнейшим продолжением и развитием настоящей работы. В целом, проведённые в работе исследования можно рекомендовать к применению в исследованиях и к других опасных заболеваний, получившим распространение в Африке в последние годы, в частности, таких, как лихорадка Эболи, от которой уже в этом 2014 г. скончалось 90 человек в соседних с Ганой Либерии и Гвинее [12].

Литература

1. *Hastie T., Tibshirani R.* Generalized Additive Models. — London: Chapman & Hall, 1990.
2. *Wood S. N.* Generalized Additive Models. — Chapman & Hall, 2006.
3. *Akimov V. A., Bykov A. A., Schetinin E. Y.* On Statistics of Extreme Values and its Applications, Monograph. — EMERCOM of Russia, FGU VNII GOCHS (FC), 2010.
4. *Banerjee S., Carlin B. P., Gelfand A. E.* Hierarchical Modelling and Analysis of Spatial Data. — Chapman & Hall / CRC, 2003.
5. *Kamman E. E., Wand M. P.* Geoadditive Models // J. Royal Stat. Soc. Series C. — 2003. — Vol. 52. — Pp. 1–18.
6. *Lang S., Brezger A.* Bayesian P-splines // J. Comp. Graph. Stat. — 2004. — No 13183–212.
7. *Eilers P., B. M.* Flexible Smoothing using B-splines and Penalties // Stat. Sci. — 1996. — Vol. 11. — Pp. 89–121.
8. *Rue H., Held L.* Gaussian Markov Random Fields, Theory and Applications. — Chapman & Hall, 2005.
9. *Fahrmeir L., Lang S.* Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors // Applied Statistics. — 2001. — Vol. 50. — Pp. 201–220.
10. *Lawson A. B.* Statistical Methods in Spatial Epidemiology. — Chichester Wiley, 2001.
11. *Borroto R. J., Martinez-Piedra R.* Geographical Patterns of Cholera in Mexico, 1991–1996 // Int. J. Epid. — 2000. — Vol. 29. — Pp. 764–772.
12. Вспышка лихорадки Эбола перекинулась на Гану. — 2014. — `http://www.rbc.ru/rbcfreenews/20140407062933.shtml`.

UDC 519.6

On Generalized Mixed-Additive Regression Models with Spatially Structural Risk Factors

Eu. Yu. Shchetinin

Department of Applied Mathematics
Moscow State Technology University "STANKIN"
3a, Vadkovsky lane, Moscow, Russian Federation, 127055

An identifying of associated risk factors which enhance the risk of infection is the most intensively growing field of epidemiology. But too little investigations considered the spatial structure of the data, as well as possible nonlinear effects of the risk factors. We developed a bayesian spatial semi-parametric regression model for cholera epidemic data. Model estimation and inference is based on fully Bayesian approach via Markov Chain Monte Carlo (MCMC) simulations. The model is applied to cholera epidemic data from Ghana, Africa. Proximity to refuse dumps, density of refuse dumps, and proximity to potential cholera reservoirs were modeled as continuous functions; presence of slum settlers and population density were modeled as fixed effects, spatial references to the communities were modeled as structured and unstructured spatial effects. We found out that the risk of cholera is associated with slum settlements and high population density. The risk of cholera is equal and lower for communities with fewer refuse dumps, but variable and higher for communities with more refuse dumps. The risk is also lower for communities distant from refuse dumps and potential cholera reservoirs. The results also indicate distinct spatial variation in the risk of cholera infection.

Key words and phrases: epidemiologic risk, bayesian regression, cholera, refuse dumps, slums, Markov chain Monte Carlo, random fields, splines.

References

1. T. Hastie, R. Tibshirani, Generalized Additive Models, Chapman & Hall, London, 1990.
2. S. N. Wood, Generalized Additive Models, Chapman & Hall, 2006.
3. V. A. Akimov, A. A. Bykov, E. Y. Schetinin, On Statistics of Extreme Values and its Applications, Monograph, EMERCOM of Russia, FGU VNII GOCHS (FC), 2010.
4. S. Banerjee, B. P. Carlin, A. E. Gelfand, PHierarchical Modelling and Analysis of Spatial Data.
5. E. E. Kamman, M. P. Wand, Geoadditive Models, J. Royal Stat. Soc. Series C 52 (2003) 1–18.
6. S. Lang, A. Brezger, Bayesian P-splines, J. Comp. Graph. Stat. (13183–212).
7. P. Eilers, M. B., Flexible Smoothing using B-splines and Penalties, Stat. Sci. 11 (1996) 89–121.
8. H. Rue, L. Held, Gaussian Markov Random Fields, Theory and Applications, Chapman & Hall, 2005.
9. L. Fahrmeir, S. Lang, Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors, Applied Statistics 50 (2001) 201–220.
10. A. B. Lawson, Statistical Methods in Spatial Epidemiology, Chichester Wiley, 2001.
11. R. J. Borroto, R. Martinez-Piedra, Geographical Patterns of Cholera in Mexico, 1991–1996, Int. J. Epid. 29 (2000) 764–772.
12. Ebola Outbreak Spread to Ghana, in Russian (2014).
URL `http://www.rbc.ru/rbcfreenews/20140407062933.shtml`