**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2016 Т. 8 № 4 С. 707–716**

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 330.4, 51-77

Распространение языков в КНР на уровне провинций: оценивание при неполных данных

Д. В. Давыдов[1,3,a], А. Б. Шаповал[2,3,b], А. И. Ямилов[2,3,c]

[1]Российская экономическая школа, Россия, 117418, г. Москва, Нахимовский пр., д. 47
[2]НИУ Высшая школа экономики, Россия, 101000, г. Москва, ул. Мясницкая, д. 20
[3]Московский государственный университет, Россия, 119991, г. Москва, ГСП-1, Ленинские горы, д. 1

E-mail: [a] dvdavuidov@econ.msu.ru, [b] abshapoval@gmail.com, [c] aibulatyamilov@gmail.com

Получено 27.06.2016, после доработки — 28.07.2016.
Принято к публикации 29.07.2016.

Данная работа посвящена решению практической задачи восстановления данных по распространению языков на региональном уровне на примере Китайской Народной Республики. Необходимость получения таких данных связана с задачей вычисления индексов лингвистического разнообразия, которые, в свою очередь, активно используются при эмпирическом анализе и прогнозе факторов социально-экономического развития, а также могут служить индикаторами потенциальных конфликтов на рассматриваемых территориях. В качестве исходной информации мы используем сведения из базы данных «Этнолог» (Ethnologue), дополняя их общедоступными данными переписей населения. Рассматриваемые нами данные содержат по каждому языку (а) оценку количества жителей страны, считающих этот язык родным, и (б) индикаторы наличия таких жителей в каждой из провинций КНР. Наша задача — для всех пар «язык–провинция» оценить количество жителей провинции, считающих этот язык родным. Она сводится к решению недоопределенной системы алгебраических уравнений. Специфика данных Ethnologue заключается в том, что, в силу большой трудоемкости и стоимости сбора таких данных, а также неполноты сведений по соответствующему разделу в переписях населения, имеющаяся информация по отдельным языкам в различных провинциях представлена за различные периоды времени. Одновременное использование таких данных приводит к тому, что возникающая система уравнений имеет неточно определенную правую часть, поэтому мы строим приближенное решение, характеризуемое минимальной невязкой. Учитывая неоднородность исходных данных (некоторые из языков оказываются на порядки менее распространенными), мы переходим к использованию взвешенной невязки, определяя в каждом уравнении весовые коэффициенты как величины, обратно пропорциональные правой части. Такой способ формирования невязки позволяет восстановить искомые переменные. Более 92 % переменных оказываются устойчивыми к изменениям правой части при вероятностном моделировании ошибок записей в исходных данных.

Ключевые слова: использование языков в регионах, индексы неоднородности, восстановление неполных данных

Исследование выполнено за счет гранта Российского научного фонда (проект № 15-18-00098). Вклад соавтора А. И. Ямилова, не входящего в научный коллектив проекта, состоит в проведении вычислительных процедур с использованием современных программных средств.

COMPUTER RESEARCH AND MODELING
2016 VOL. 8 NO. 4 P. 707–716

Ки&М

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 330.4, 51-77

Languages in China provinces: quantitative estimation with incomplete data

D. V. Davydov[1,3,a], A. B. Shapoval[2,3,b], A. I. Yamilov[2,3,c]

[1]New Economic School, 47 Nakhimovskii Prospekt, 117418, Moscow, Russia
[2]NRU Higher School of Economics, 20 Myasntiskaya st., Moscow, 101000, Russia
[3]Lomonosov Moscow State University, 1 Leninskie gory, Moscow, 119991, Russia

E-mail: [a] dvdavuidov@econ.msu.ru, [b] abshapoval@gmail.com, [c] aibulatyamilov@gmail.com

Received 27.06.2016, after completion — 28.07.2016.
Accepted for publication 29.07.2016.

This paper formulates and solves a practical problem of data recovery regarding the distribution of languages on regional level in context of China. The necessity of this recovery is related to the problem of the determination of the linguistic diversity indices, which, in turn, are used to analyze empirically and to predict sources of social and economic development as well as to indicate potential conflicts at regional level. We use Ethnologue database and China census as the initial data sources. For every language spoken in China, the data contains (a) an estimate of China residents who claim this language to be their mother tongue, and (b) indicators of the presence of such residents in China provinces. For each pair language/province, we aim to estimate the number of the province inhabitants that claim the language to be their mother tongue. This base problem is reduced to solving an undetermined system of algebraic equations. Given additional restriction that Ethnologue database introduces data collected at different time moments because of gaps in Ethnologue language surveys and accompanying data collection expenses, we relate those data to a single time moment, that turns the initial task to an 'ill-posed' system of algebraic equations with imprecisely determined right hand side. Therefore, we are looking for an approximate solution characterized by a minimal discrepancy of the system. Since some languages are much less distributed than the others, we minimize the weighted discrepancy, introducing weights that are inverse to the right hand side elements of the equations. This definition of discrepancy allows to recover the required variables. More than 92% of the recovered variables are robust to probabilistic modelling procedure for potential errors in initial data.

Keywords: regional languages usage, dissimilarity indices, incomplete data identification

Citation: *Computer Research and Modeling*, 2016, vol. 8, no. 4, pp. 707–716 (Russian).

The authors wish to thank the Russian Science Foundation for its financial support through the research project 15-18-00098.

Введение

Лингвистическое разнообразие как предмет изучения имеет достаточно длительную историю. Общий обзор, отражающий особенности исследования разнообразия, можно найти, например, в [Вебер, Давыдов, 2015a]. Теоретические подходы к построению количественных оценок неоднородности распространения языков известны со второй половины XX века [Greenberg, 1956] и продолжают активно развиваться [Акчурина и др., 2015; Ginsburgh, Weber, 2011]. Индексы этнического, лингвистического и религиозного разнообразия активно используются в эмпирических моделях анализа факторов социально-экономического развития [Alesina et al., 2003; Florida, 2002; Mauro, 1995; Montalvo, Reynal-Querol, 2003], а также могут служить вспомогательными индикаторами потенциальных конфликтов на рассматриваемых территориях [Easterly, Levine, 1997; Esteban, Mayoral, Ray, 2012; Fearon, Laitin, 2003; Montalvo, Reynal-Querol, 2005; Бутаева, Вебер, Давыдов, 2016; Вебер, Давыдов, Довер, 2015]. Несмотря на это, существенная часть эмпирических исследований влияния языковой и этнической неоднородности на экономическое развитие и конфликты посвящена межстрановому анализу (см., например, [Alesina, Zhuravskaya, 2011; Fearon, 2003]) и не затрагивает аспекты неоднородности внутри страны. Это связано прежде всего с проблемами сбора данных о неоднородности на уровне отдельных регионов (штатов, провинций, округов и т. п.), а также во многом объясняется высоким уровнем сопутствующих издержек. В то же время переход от межстранового уровня описания и анализа неоднородности к аналогичному анализу внутри стран позволяет существенно более точно характеризовать причины и условия соответствующих региональных различий и, как правило, демонстрирует более высокую статистическую значимость индексов неоднородности [Вебер, Давыдов, 2015b]. Переход на региональный уровень анализа в отношении лингвистического разнообразия до некоторой степени возможен при использовании широко известной для профессионалов базы данных «Этнолог» [Ethnologue, 2009], где указаны сведения о распространении языков на субстрановом уровне. Однако эти данные не являются достаточными для получения адекватных количественных оценок, а именно вычисления большинства индексов неоднородности, так как содержат только индикатор наличия каждого языка в каждом из рассматриваемых регионов, но не количество людей в регионе, владеющих соответствующими языками. В данной статье мы рассматриваем практическую задачу восстановления данных по распространению языков на региональном уровне на примере КНР. В качестве исходной информации мы используем сведения из базы данных «Этнолог» [Ethnologue, 2009] по распространению языков в провинциях КНР, а также по общему числу носителей языков в целом по стране, дополняя их данными переписи населения КНР. Конечной задачей является получение оценки численности жителей по каждому языку в каждой провинции. В силу большой трудоемкости сбора соответствующих данных имеющаяся информация по отдельным языкам в различных провинциях представлена в базе данных «Этнолог» [Ethnologue, 2009] за различные периоды времени. Совмещение таких разнородных по времени данных формально приводит к не точно определенной системе алгебраических уравнений, поиску адекватного приближенного решения которой и посвящена основная часть данной работы.

Описание данных и формальная постановка задачи

Исходя из структуры данных переписи населения за 2012 год, на территории КНР выделена 31 провинция, по каждой из которых известны данные о количестве P_i проживающих в ней людей ($i \in I = \{1, \ldots, 31\}$). Согласно базе данных «Этнолог» [Ethnologue, 2009] в КНР распространены 296 языков. По каждому языку $j \in J = \{1, \ldots, 296\}$ доступны сведения о количестве Q_j людей, проживающих в целом на территории КНР, которые считают данный язык родным. При этом предполагается, что люди считают родным только один язык. Кроме того, для

каждой провинции i известно подмножество J_i языков, распространенных на территории данной провинции. Иначе говоря, для всех $i \in I$ и $j \in J \setminus J_i$ значения $x_{ij} \equiv 0$, где через x_{ij} обозначено количество людей, которые проживают в провинции i и считают родным язык j. Из исходных данных [Ethnologue, 2009] следует, что только 440 неизвестных x_{ij} отличны от нуля.

Переменные x_{ij} удовлетворяют двум видам ограничений. Во-первых, их сумма по второму индексу при фиксированном первом определяет известное число жителей i-й провинции:

$$\sum_{j \in J} x_{ij} = \sum_{j \in J_i} x_{ij} = P_i, \quad i = 1, \ldots, 31. \tag{1}$$

Аналогично: суммирование x_{ij} по первому индексу при фиксированном втором задает количество людей, у которых j-й язык является родным:

$$\sum_{i \in I} x_{ij} = Q_j, \quad j = 1, \ldots, 296. \tag{2}$$

Совокупность условий (1), (2) определяет систему из 327 линейных уравнений и по структуре повторяет ограничения однопродуктовой транспортной задачи, что гарантирует ее разрешимость в условиях сбалансированности (см., например, [Александрова и др., 2013]):

$$\sum_i P_i = \sum_j Q_j.$$

Однако особенностью данной системы является ее некорректная постановка в силу неточно определенной правой части: сумма всех проживающих в КНР ($P_1 + \ldots + P_{31}$), согласно имеющимся эмпирическим данным, не равна сумме всех говорящих на языках, распространенных на территории КНР ($Q_1 + \ldots + Q_{296}$), хотя по своему определению это одно и то же число — количество жителей КНР. Причины такого расхождения в данных описаны выше, во введении. В результате решение системы уравнений (1), (2) приходится искать приближенно.

Для формулировки приближенной задачи удобно ввести новые обозначения, чтобы система уравнений записывалась в матричном виде:

$$Ay = b, \tag{3}$$

где A — это матрица коэффициентов, y — вектор, содержащий неизвестные, $b > 0$ — вектор правой части. Координатами вектора b последовательно являются положительные значения $P_1, \ldots, P_{31}$, $Q_1, \ldots, Q_{296}$. Вектор y получается корректным переименованием не равных нулю неизвестных x_{ij} и содержит 440 компонент. Матрица A, в соответствии с уравнениями (1), (2), содержит только нулевые либо единичные элементы.

Поиск приближенного решения

С целью нахождения наилучшего приближенного решения задачи мы используем метод минимизации относительной невязки. Для системы (3) зададим вектор невязок Δ с компонентами

$$\Delta_k = \frac{\sum_{l=1}^{440} a_{kl} y_l - b_k}{b_k}, \quad k = 1, \ldots, 327,$$

и сформулируем задачу минимизации

$$\Delta = \sum_{k=1}^{327} \Delta_k^2 \longrightarrow \min \tag{4}$$

при условии неотрицательности

$$y_l \geq 0, \quad l = 1, \ldots, 440. \tag{5}$$

Вопрос выбора «наилучшего» минимизирущего функционала хорошо известен: необходимо стремиться, чтобы на найденном решении вклад всех уравнений в оптимальную невязку был сбалансирован.

Минимизация абсолютной невязки (суммы квадратов разностей между левой и правой частями уравнения) для решаемой приближенно системы (3) приводит к неустойчивости решения относительно малых изменений правых частей. Попутно также отметим, что стандартный SVD-метод (разложение по сингулярным числам матрицы) поиска приближенного решения системы (3) не подразумевает выполнения условий неотрицательности (5), и его применение приводит к наличию отрицательных координат в векторе решения. Тем не менее решение системы (3) SVD-методом дает в определенном смысле «оптимальную» абсолютную невязку $4.4 \cdot 10^6$, с которой естественно сравнивать невязку, полученную другими методами.

Оптимизационная задача (4), (5) решена с помощью функции `quadprog` в математическом пакете *Matlab*. Абсолютная и относительная невязки этого решения равны $4.5 \cdot 10^8$ и 3.18 соответственно. Заметим, что абсолютная невязка выросла на два порядка по сравнению с (заведомо неправильным) решением, полученным SVD-методом и содержащим отрицательные координаты.

Учитывая высокую разреженность соответствующих данных, их общее агрегированное представление в табличном виде является неэффективным. В то же время интерес представляют результаты расчетов для относительно представительных языков, локализованных в одной из провинций, а также распространение по провинциям наиболее представительных диалектов китайского языка.

В первом случае мы выделяем порог отсечения в размере 1 млн носителей соответствующего языка. По нашим оценкам, носители монгольского языка преимущественно сосредоточены в провинции Хэйлунцзян (свыше 3.5 млн чел.); носители корейского языка, соответственно, в провинции Гирин (чуть менее 3 млн чел.), а в Синьцзян-Уйгурском автономном районе преимущественно представлены казахский (1.3 млн чел.) и уйгурский (9.8 млн чел.) языки.

Во втором случае мы выделяем порог отсечения в размере 10 млн носителей соответствующего диалекта китайского языка, что позволяет представить результаты наших расчетов в виде таблицы 1.

Далее мы оцениваем устойчивость полученного решения относительно изменения исходных данных.

Устойчивость решения относительно исходных данных

Для оценки полученного «приближенного» решения задачи (3) в форме решения оптимизационной задачи (4), (5) воспользуемся следующей вероятностной схемой варьирования данных.

Будем считать, что при записи информации возможны ошибки. Предположим, что регистрация человека, считающего язык j своим родным языком, происходит следующим образом. Пусть α_j — это некоторое число из отрезка $[0, 1]$. С вероятностью $1 - \alpha_j$ регистрация данного человека происходит корректно: записано, что он считает язык j родным. С вероятностью α_j регистрация ошибочна. Тогда его соотносят с другим языком $\hat{j}$. Предполагается, что это событие имеет вероятность $\beta_{\hat{j}}$, $\hat{j} \in J$. Мы полагаем, что

$$\alpha_j = \frac{A}{\sqrt{Q_j}}, \quad \beta_{\hat{j}} = \frac{\sqrt{Q_{\hat{j}}}}{\sum_{j=1}^{296} \sqrt{Q_j}}, \ \hat{j} \neq j, \hat{j} \in J,$$

Таблица 1. Оценка распространения основных диалектов китайского языка по провинциям КНР (млн чел.)

Провинции и автономные округа КНР	Диалекты китайского языка							
	Gan	Hakka	Jinyu	Mandarin	Min Bei	Min Nan	Wu	Yue
Anhui	0	0	0	**61**	0	0	0	0
Beijing	0	0	**11**	0	0	0	0	0
Chongqing	0	0	0	**61**	0	0	0	0
Fujian	**8**	**6**	0	0	**2**	0	**4**	0
Gansu	0	0	0	**25**	0	0	0	0
Guangdong	0	**10**	0	0	0	**5**	0	**55**
Guangxi	0	0	0	**29**	0	0	0	0
Hainan	0	0	0	0	0	**2**	0	0
Hebei	0	0	**17**	0	0	0	0	0
Hubei	0	0	0	**56**	0	0	0	0
Hunan	0	0	0	**59**	0	0	0	0
Jiangsu	0	0	0	0	0	0	**63**	0
Jiangxi	**14**	**8**	0	0	**3**	**4**	**6**	0
Neimenggu	0	0	**17**	0	0	0	0	0
Qinghai	0	0	0	**4**	0	0	0	0
Shaanxi	0	0	0	**36**	0	0	0	0
Shanxi	0	0	**21**	0	0	0	0	0
Sichuan	0	0	0	**81**	0	0	0	0
Yunnan	0	0	0	**29**	0	0	0	0
Zhejiang	0	0	0	0	**5**	**13**	**15**	0

где константа A выбрана равной 0.005 или 0.01. По нашим предположениям, чем представительнее язык, тем меньше вероятность ошибки и тем больше шансов, что носители других языков будут приписаны к данному. Разумеется, предложенная реализация данного тезиса не является единственно возможной.

В процессе проверки решения задачи (4), (5) на чувствительность к исходным данным мы убедились, что качественный характер результатов сохраняется при различных постулируемых зависимостях вероятностей α и β от представительности языка.

Пусть S_k — это случайная величина, которая означает количество людей, «приписанных» к языку $(k-31)$, $k = 32, \ldots, 327$. Значения b_k в системе (3), $k = 32, \ldots, 327$, можно рассматривать как точечные оценки среднего значения соответствующих случайных величин S_k. В предположении о независимости ошибок регистрации данных можно показать, что дисперсия относительной ошибки S_k/b_k при идентификации значения S_k не превосходит

$$\sigma_k^2 = \left(2A\sqrt{\tilde{Q}_{k-31}} - A^2 - \frac{A\tilde{Q}_{k-31}}{\varkappa} - 295\frac{A^2\tilde{Q}_{k-31}}{\varkappa^2}\right)\Big/ b_k^2, \tag{6}$$

где

$$\tilde{Q}_{k-31} = \frac{b_k}{1 - A/\varkappa}, \quad \varkappa = \sum_{i=1}^{296} \sqrt{\tilde{Q}_i}.$$

Здесь $\tilde{Q}_k$ — это оценка количества говорящих на языке k, $k = 1, \ldots, 296$.

Для оценки устойчивости найденного решения предложена и осуществлена следующая процедура. $N = 100$ раз генерируется значение нормальной случайной величины $\zeta_k \sim \mathbf{N}(0, \sigma_k^2)$ с нулевым математическим ожиданием и дисперсией σ_k^2, $k = 32, \ldots, 327$. В векторе b правых частей системы (3) координаты $b_1, \ldots, b_{31}$, описывающие принадлежность людей к провинциям,

считаются достоверными и потому сохраняются неизменными при случайной вариации данных: $\tilde{b}_k = b_k, \quad k = 1, \ldots, 31$. Координаты $b_{32}, \ldots, b_{327}$ заменяются на соответствующие случайные реализации $\tilde{b}_k = b_k e^{\zeta_k}$. Далее решается задача оптимизации (4), (5), где в (4) для всех $k = 1, \ldots, 327$ произведена замена b_k на $\tilde{b}_k$. Полученные N векторов решений упорядочиваются по возрастанию в соответствии с оптимальным значением относительной невязки Δ, определенной в (4). В качестве тестового вектора $\tilde{y}$ выбрано решение, у которого относительная невязка Δ составляет верхний 5%-й квантиль во множестве относительных невязок (иными словами, доля относительных невязок, превосходящих данную, равна 0.05). Затем вычисляются относительные отклонения исходного решения по отношению к тестовому $\tilde{y}$:

$$d_l = \frac{|\tilde{y}_l - y_l|}{y_l}, \quad l = 1, \ldots, 440,$$

для всех 440 координат решения y.

Функция распределения относительных отклонений d_l приведена на рисунке 1.

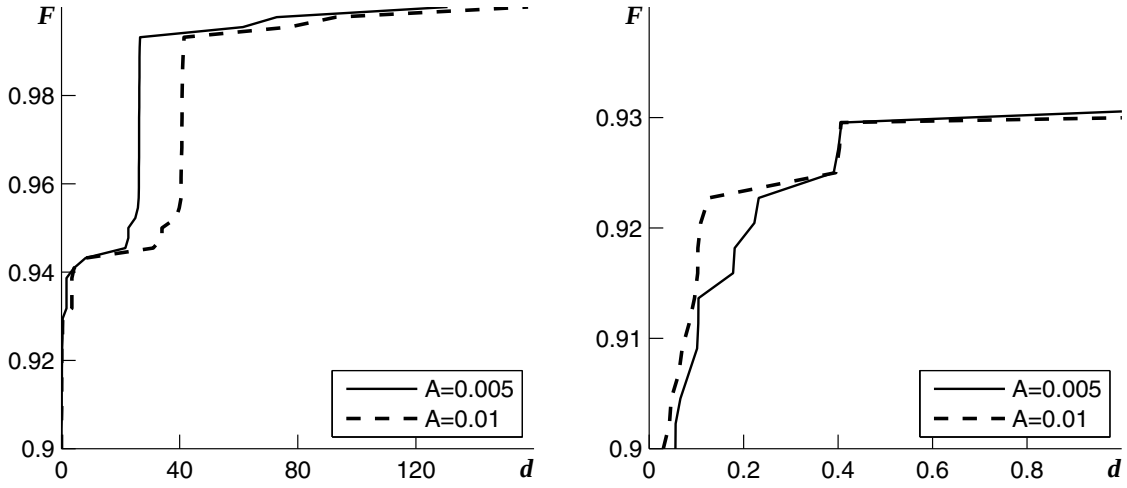

Рис. 1. Функция распределения относительных отклонений d_l, полученных в тестовом решении с вероятностью ошибок записи $A/\sqrt{Q_j}$, где A выбрано равным 0.005 и 0.01 (полный график — слева; его часть в укрупненном масштабе — справа)

Структура графика, полученного при различных значениях A, показывает, что более 92 % полученных значений остаются устойчивыми относительно введенных возмущений. Поэтому значения 405 переменных из 440 мы считаем корректно восстановленными. Значения остальных 35 переменных мы рассматриваем как неробастные оценки для идентифицируемых нами данных.

Переменные, робастность восстановления которых не подтверждена, преимущественно находятся среди тех координат решения y, значения которых малы. Этот результат ожидаем. Используемая нами процедура минимизации суммарной относительной невязки Δ в (4) предназначена для выравнивания относительных невязок Δ_k, вычисленных по полученному решению. Структура данных, однако, такова, что наряду с этническими группами численностью в миллионы людей в КНР представлены десятки этнических групп, содержащих несколько десятков человек. Поэтому выравнивание относительных невязок происходит лишь частично. В результате значительные относительные изменения некоторых координат вектора-решения с малыми номинальными значениями несущественно влияет на суммарную относительную невязку Δ.

Иллюстрируя местоположение плохо восстановленных переменных, мы упорядочили координаты вектора-решения по возрастанию и построили функцию распределения F полученной

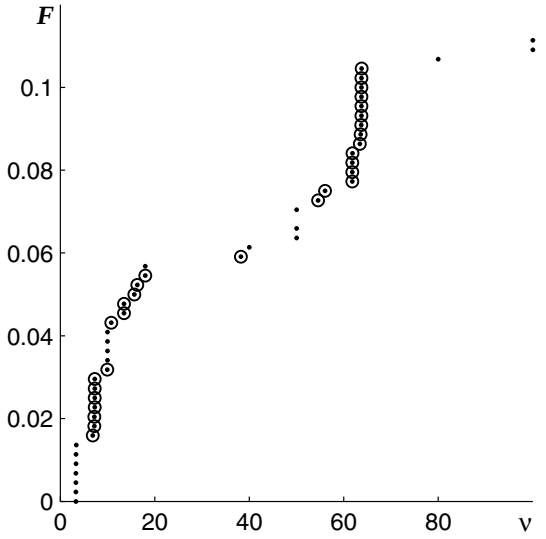

Рис. 2. Левый фрагмент функции распределения координат вектора решений $y = (y_1, \ldots, y_{327})$, упорядоченных в порядке возрастания (черные точки); кругами выделены переменные, корректность восстановления которых сомнительна; $A = 0.005$

перестановки. Другими словами, $F(\eta)$ — это доля координат вектора y, значения которых не превосходят η. Левый фрагмент данной функции (координаты не превосходят 100) представлен на рисунке 2 черными точками. Кругами выделены переменные, робастность восстановления которых не подтверждается нашим тестом. За пределами фрагмента, представленного на рисунке 2, находится только 5 неробастных переменных (со значениями 248, 1144, 456124 и 519635, округлено до целых). Мы заключаем, что переменные, не обладающие свойством робастности, как правило, малы по значению. Следовательно, применяемая нами идея по минимизации относительной (а не абсолютной) невязки, которая в целом демонстрирует хороший результат, оказывается недостаточно эффективной для переменных, значения которых составляют 1–2 порядка магнитуд.

Заключение

Рассмотренное на примере КНР решение задачи идентификации данных о региональном распределении лингвистически неоднородного населения имеет как теоретические, так и практические аспекты.

С практической точки зрения необходимость выполнения предложенных в работе вычислительных процедур обусловлена дальнейшим использованием полученных данных о численности носителей различных языков на уровне провинций КНР в эмпирическом (эконометрическом) анализе социально-экономического развития и влияния неоднородности на возникновение конфликтов.

В общетеоретическом контексте в работе предложены и апробированы алгоритм восстановления неполных данных, а также подходы к оценке робастности получаемых значений численности носителей языков в различных регионах страны. Демонстрируемые нами количественные результаты на примере КНР позволяют ожидать достаточно высокую точность и устойчивость вычислений при применении указанного алгоритма к исходным данным базы данных «Этнолог» по другим странам. При этом, несомненно, следует контролировать специфику исходных данных (как количество регионов страны, так и микроданные по малораспространенным языкам) и адаптировать их при необходимости для повышения общей эффективности алгоритма.

Одним из таких вариантов адаптации получаемых по предложенному алгоритму результатов для дальнейших эконометрических расчетов является пересмотр исходных данных с исключением из рассмотрения локальных малораспространенных языков с малым числом носителей в целом по стране и соответствующим (пропорциональным) уточнением общей численности населения. Мы предполагаем, что подобная коррекция данных, с одной — не приводит к возникновению существенных погрешностей в последующем прикладном эмпирическом анализе, а с другой стороны, увеличивает общий показатель робастности решения рассматриваемой нами системы уравнений.

Список литературы (References)

Акчурина Д. Д., Вебер Ш., Давыдов Д. В., Крутиков Д. В., Хазанов А. А. Измерение разнообразия: теория и социально-экономические приложения // Современная экономика: проблемы и решения. — 2015. — № 2. — С. 8–28.

Akchurina D. D., Weber S., Davydov D. V., Krutikov D. V., Khazanov A. A. Izmerenie raznoobraziya: teoriya i sotsialno-ekonomicheskie prilozheniya [Measuring diversity: theory and socio-economic applications] // Sovremennaya ekonomika: problemy i resheniya. — 2015. — No. 2. — S. 8–28 (in Russian).

Александрова И. А. и др. Методы оптимальных решений в экономике и финансах: Учебник / Под ред. В. М. Гончаренко, В. Ю. Попова. — М: КНОРУС, 2013.

Aleksandrova I. A. et al. Metody optimalnyh resheniy v ekonomike i finansah: Uchebnik [Methods for optimal solutions in economics and finance] / Pod red. V. M. Goncharenko, V. Yu. Popova. Moskva: KNORUS, 2013 (in Russian).

Бутаева К. О., Вебер Ш., Давыдов Д. В. Язык, культура, миграция, конфликты: экономическая проекция // Вестник Московского университета. Сер. 6: Экономика. — 2016. — № 1. — С. 3–21.

Butaeva K. O., Weber S., Davydov D. V. Yazyk, kultura, migratsiya, konflikty: ekonomicheskaya proyektsiya [Language, culture, migration, and conflicts: economic view] // Vestnik Moskovskogo Universiteta. Ser. 6: Ekonomika. — 2016. — No. 1. — S. 3–21 (in Russian).

Вебер Ш., Давыдов Д. В. Экономика разнообразия: подходы, методы, результаты // Экономика и математические методы. — 2015. — Т. 51, № 4. — С. 3–13.

Weber S., Davydov D. V. Ekonomika raznoobraziya: podhody, metody, rezultaty [Economics of diversity: approaches, methods, results] // Ekonomika i matematicheskie metody. — 2015. — T. 51, No. 4. — S. 3–13 (in Russian).

Вебер Ш., Давыдов Д. В. Этнокультурные факторы, социальные процессы, конфликты и экономическое развитие // Современная экономика: проблемы и решения. — 2015. — № 8. — С. 129–143.

Weber S., Davydov D. V. Etnokulturnye faktory, sotsialnye protsessy, konflikty i ekonomicheskoye razvitie [Ethnocultural factors, social processes, conflicts, and economic development] // Sovremennaya ekonomika: problemy i resheniya. — 2015. — No. 8. — S. 129–143 (in Russian).

Вебер Ш., Давыдов Д. В., Довер П. А. Трансферты и предотвращение конфликтов: «За» и «Против» // Экономика и математические методы. — 2015. — Т. 51, № 2. — С. 60–69.

Weber S., Davydov D. V., Dover P. A. Transferty i predotvrashchenie konfliktov: «Za» i «Protiv» // Ekonomika i matematicheskie metody. — 2015. — T. 51, No. 2. — S. 60–69.

Alesina A., Devleeschouwer A., Easterly W., Kurlat S., Wacziarg R. Fractionalization // Journal of Economic Growth. — 2003. — Vol. 8, No. 2. — P. 155–194.

Alesina A., Zhuravskaya E. Segregation and the Quality of Government in a Cross-section of Countries // American Economic Review. — 2011. — Vol. 101, No. 5. — P. 1872–1911.

Easterly W., Levine R. Africa's growth tragedy: Policies and ethnic divisions // The Quarterly Journal of Economics. — 1997. — Vol. 112, No. 4. — P. 1203–1250.

Esteban J., Mayoral L., Ray D. Ethnicity and Conflict: An Empirical Study // American Economic Review. — 2012. — Vol. 102, No. 4. — P. 1310–1342.

Ethnologue, Languages of the World. M. Paul Lewis, ed. Dallas, TX: SIL International. 2009.

Fearon J. Ethnic and Cultural Diversity by Country // Journal of Economic Growth. — 2003. — Vol. 8, No. 2. — P. 195–222.

Fearon J. D., Laitin D. D. Ethnicity, Insurgency, and Civil War // American Political Science Review. — 2003. — Vol. 97, No. 1. — P. 75–90.

Florida R. The Rise of the Creative Class. New York: Basic Books, 2002.

Greenberg J. The measurement of linguistic diversity // Language. — 1956. — Vol. 32. — P. 109–115.

Ginsburgh V., Weber S. How Many Languages Do We Need? The Economics of Linguistic Diversity. — Princeton, NJ: Princeton University Press, 2011.

Mauro P. Corruption and Growth // The Quarterly Journal of Economics. — 1995. — Vol. 110, No. 3. — P. 681–712.

Montalvo J. G., Reynal-Querol M. Religious Polarization and Economic Development // Economics Letters. — 2003. — Vol. 80, No. 2. — P. 201–210.

Montalvo J. G., Reynal-Querol M. Ethnic Polarization, Potential Conflict and Civil Wars // American Economic Review. — 2005. — Vol. 95, No. 3. — P. 796–816.