УДК ОО4+658

Хубаев Георгий Николаевич
доктор экономических наук, профессор,
профессор кафедры информационных систем и прикладной информатики,
Ростовский государственный экономический университет
gkhubaev@mail.ru
Georgy N. Khubaev
dr of Economics, professor, Rostov State University of Economics
gkhubaev@mail.ru

МЕТОД ВЫДЕЛЕНИЯ ИСКОМОГО ПОДМНОЖЕСТВА ОБЪЕКТОВ ИЗ МНОЖЕСТВА БОЛЬШОЙ МОЩНОСТИ[1]

THE METHOD OF SELECTING THE DESIRED SUBSET OF OBJECTS FROM A SET OF SUFFICIENTLY LARGE POWER IS PROPOSED AND SUCCESSFULLY TESTED

Аннотация. *Предложен и успешно апробирован оригинальный метод выделения искомого подмножества объектов из множества достаточно большой мощности. Использование метода для оценки и прогнозирования показателей уровня жизни населения субъектов РФ позволило построить совокупность регрессионных моделей *лучшего качества (при оценке по стандартным критериям статистической значимости - $R^2_{скорр}>0,9$; $F_{кр}>>100$, в большинстве случаев у b–коэффициентов отношение $b_i/Б_{bi}>>2$); *с лучшими прогнозными свойствами; *с использованием открытых официальных статистических данных и одновременно с проверкой на наличие аномальных наблюдений в массиве исходной информации, содержащем десятки тысяч числовых данных.*

Ключевые слова: *уровень жизни населения, субъекты РФ, регрессионные модели, статистическая значимость*

Abstract. *The original method of selecting the desired subset of objects from a set of sufficiently large power is proposed and successfully tested. The use of the method for the assessment and forecasting of indicators of living standards of the population of subjects of the Russian Federation has allowed to build a set of regression models *better quality (in the assessment according to standard criteria of statistical significance - R2скорр>0,9; Fкр >>100, in most cases, the b-coefficients the ratio bi/Бbi >>2); *with the best predictive properties; *using open official*

[1] *Статья подготовлена по результатам исследований, выполненных при поддержке Российского фонда фундаментальных исследований (РФФИ)* – проект **18-010-00806/18 «УРОВЕНЬ ЖИЗНИ НАСЕЛЕНИЯ АДМИНИСТРАТИВНО-ТЕРРИТОРИАЛЬНЫХ ОБРАЗОВАНИЙ: выявление, исследование, анализ и оценка значимости определяющих факторов (для последующей оптимизации в условиях ограниченных ресурсов)»**

statistics and at the same time checking for anomalous observations in an array of initial information containing tens of thousands of numerical data.

Key words: *standard of living, subjects of the Russian Federation, regression models, statistical significance*

ПОСТАНОВКА ЗАДАЧИ. Известно, что при проведении экспертиз по упорядочению объектов не рекомендуется предлагать эксперту для ранжирования более 15-16 объектов. При большем количестве объектов следует разбивать их на группы и выполнять ранжирования отдельно для каждой группы объектов.

Очевидно, что в этом случае возникает ряд неудобств. Во-первых, группировки можно осуществить бесчисленным количеством способов. И, во-вторых, как определить, в каких группах оказались искомые (самые важные, нужные, полезные, …) объекты. В то же время при исследовании, например, уровня жизни населения субъектов РФ или уровня жизни населения в какой-либо стране мира приходится анализировать влияние на интересующий исследователя показатель сотен и тысяч различных факторов. Так, в открытых статистических данных, сформированных ООН и Всемирным банком, представлены сведения о значениях нескольких тысяч социально-экономических показателей, характеризующих уровень социально-экономического развития стран мира, а в сборниках Росстата – сведения о сотнях показателей уровня социально-экономического развития субъектов РФ.

Спрашивается, какой показатель из этого множества оказывает наиболее существенное, определяющее влияние на уровень жизни населения?

Как выделить минимальную по составу группу показателей для последующего построения прогнозных и/или оптимизационных моделей?

Ведь в настоящее время отсутствуют и в России, и за рубежом корректные, обоснованные методы и/или программно реализованные алгоритмы, позволяющие оперативно осуществить ранжирования сотен и тысяч объектов по критерию, заданному исследователем.

В статье предложен и успешно апробирован оригинальный метод выделения ограниченного подмножества объектов (факторов, признаков, показателей) из исходного множества достаточно большой мощности, содержащего сотни и тысячи объектов.

1.Выделение ограниченного подмножества искомых объектов-факторов. Предлагаемая последовательность шагов:

Шаг 1. С использованием таблицы или датчика случайных чисел из базы данных, содержащей сведения об экспертах, возможных участниках различных экспертиз, выбираются компетентные в исследуемой предметной области потенциальные участники экспертизы по выявлению и последующему упорядочению объектов-факторов, оказывающих наиболее существенное влияние на изучаемый показатель.

Шаг 2. Выбранным экспертам предлагают принять участие в решении конкретной, интересующей организаторов экспертного опроса задачи.

Шаг 3. Экспертам, согласившимся участвовать в опросе, присваивают идентификаторы (также с использованием датчика случайных чисел). Предположим, что согласившихся участвовать в экспертизе оказалось 100.

[**Замечание 1.** Все три шага и ряд последующих шагов выполняются автоматически, т.е. не только участники, но и организаторы экспертизы не знают, кто конкретно участвует в опросах, кто и как обосновал своё решение, как возникают группировки участников опроса]

Шаг 4. Каждого участника экспертного опроса информируют о необходимости перечислить факторы-показатели, оказывающие, по мнению эксперта, основное влияние на исследуемый показатель. и выполнить ранжирования (упорядочение) перечисленных факторов по степени значимости.

[**Замечание 2.** Необходимость выполнять не только выбор значимых факторов, но и их ранжирования, вынуждает большинство экспертов более внимательно отнестись к выбору подмножества искомых факторов]

Шаг 5. У каждого эксперта в списках выделенных и упорядоченных факторов оставляют не более 15-16 факторов.

В результате выполнения шага 5 будет сформирована таблица вида таблицы 1.

[**Замечание 3.** Экспертов, у которых в списках оказались факторы, выбранные не более, чем 10-15 процентами участвующих в экспертизе – в таблице 1 это факторы X_2, X_j, X_{j+k} - просят объяснить причины выбора именно этих факторов, и с объяснениями знакомят всех экспертов, предлагая при желании изменить свои ранжирования].

Таблица 1 - Результаты экспертизы по формированию ограниченного подмножества существенных факторов

Эксперт	**По мнению участников опроса, именно эти факторы оказывают основное влияние на исследуемый показатель**								
	X_1	X_2	…	X_j	X_{j+1}	…	X_{j+k}	…	X_m
Z_1	1	1	…	1	1	…	0	…	1
Z_2	0	0	…	0	1	…	1	…	0
Z_3	1	1	…	1	1	…	0	…	1
…	…	…	…	…	…	…	…	…	…
Z_i	0	1	…	0	0	…	0	…	0
…	…	…	…	…	…	…	…	…	…
$\sum X_j$	92	11	…	3	97	…	2	…	95

Шаг 6. Обработка результатов экспертизы – таблицы 1.

Пусть $Z=|z_i|$, (i=1, 2, …) – множество экспертов, которым с использованием таблицы или датчика случайных чисел присвоены идентификаторы Zi. Исходная информация представляется в виде таблицы $\{x_{ij}\}$. При этом

$$x_{ij} = \begin{cases} 1, \text{если } i-\text{й эксперт выбрал } j-\text{й фактор;} \\ 0, \text{если } j-\text{й фактор отсутствует в списке у } i-\text{го эксперта.} \end{cases}$$

Выделим экспертов Z_i и Z_k (i, k =1, 2,…) и введем следующие обозначения: $P_{ik}^{(11)}$ – число факторов, выбранных одновременно Z_i и Z_k, т.е. $P_{ik}^{(11)} = |Z_i \cap Z_k|$- мощность пересечения множеств $Z_i = \{x_{ij}\}$ и $Z_k = \{x_{kj}\}$ $(j \in \overline{1,m}; x \mid x_{ij},\ x_{ij} = 1)$; $P_{ik}^{(10)}$ - число факторов, выбранных экспертом Z_i, но отсутствующих в списке Z_k, т.е. $P_{ik}^{(10)} = |Z_i / Z_k|$- мощность разности множеств $Z_i = \{x_{ij}\}$ и $Z_k = \{x_{kj}\}$; $P_{ik}^{(01)}$ - число факторов, отсутствующих в списке Z_i, но выбранных Z_k, т.е. $P_{ik}^{(01)} = |Z_k / Z_i|$.

В качестве меры рассогласования между строками Z_i и Z_k выберем величину $S_{ik} = P_{ik}^{(01)}/(P_{ik}^{(11)} + P_{ik}^{(10)})$, а для оценки степени поглощения экспертом Z_k списка факторов эксперта Z_i (степени включения, «вхождения» списка факторов эксперта Z_i в Z_k) – величину $h_{ik} = P_{ik}^{(11)}/(P_{ik}^{(11)} + P_{ik}^{(10)})$.

Построим матрицы $P = \{p_{ik}^{(01)}\}, S = \{s_{ik}\},\ G = \{g_{ik}\}, H = \{h_{ik}\}$ $(i,k \in \overline{1,n})$, где $g_{ik} = P_{ik}^{(11)}/(P_{ik}^{(11)} + P_{ik}^{(10)} + P_{ik}^{(01)})$ – мера подобия Жаккарда.

Преобразуем P, S, G и H в логические матрицы отношения поглощения (включения) для значений $\varepsilon_p, \varepsilon_s, \varepsilon_g,\ \varepsilon_h$.

$P_0 = \{p_{ik}^0\},\ S_0 = \{s_{ik}^0\},\ G_0 = \{g_{ik}^0\},\ H_0 = \{h_{ik}^0\}$ $(i,k \in \overline{1,n})$,

элементы которых определяются следующим образом:

$$P_{ik}^0 = \begin{cases} 1, \text{если } P_{ik}^{(01)} \le \varepsilon_p \text{ и } i \ne k, \\ 0, \text{если } P_{ik}^{(01)} > \varepsilon_p \text{ или } i = k; \end{cases} \quad S_{ik}^0 = \begin{cases} 1, \text{если } S_{ik} \le \varepsilon_s \text{ и } i \ne k, \\ 0, \text{если } S_{ik} > \varepsilon_s \text{ или } i = k; \end{cases}$$

$$g_{ik}^0 = \begin{cases} 1, \text{если } g_{ik} \ge \varepsilon_g \text{ и } i \ne k, \\ 0, \text{если } g_{ik} < \varepsilon_g \text{ или } i = k; \end{cases} \quad h_{ik}^0 = \begin{cases} 1, \text{если } h_{ik} \ge \varepsilon_h \text{ и } i \ne k\ , \\ 0, \text{если } h_{ik} < \varepsilon_h \text{или } i = k, \end{cases}$$

где ε - выбранные граничные значения.

Разницу в составе факторов, выбранных участниками экспертного опроса, можно наглядно показать на графах, построенных по матрицам G_0 и H_0. Степень взаимосвязи экспертов по составу выбранных ими факторов можно оценить, анализируя матрицу $G = \{g_{ik}\}$.

Для оценки информацционного веса выбранных факторов по матрице P_0 найдем P_0^2 и сумму $(P_0 + P_0^2)$. Анализ матрицы $(P_0 + P_0^2)$ позволяет определить, какой из факторов, по мнению участников экспертного опроса, имеет наибольший информационный вес (ранг).

Метод и разработанные на его основе программные продукты позволяют оперативно проводить сравнительный анализ *практически неограниченного количества факторов и мнений экспертов*, корректно и *с минимальными* трудозатратами *осуществлять* *классификацию (группировку) экспертов в зависимости от состава выбранных факторов-показателей; *формирование полного перечня факторов, выделенных участниками экспертизы; **количественную оценку* информационного веса каждого фактора.

Шаг 7. По данным таблицы 1 формируется список из 16 факторов для дальнейшего исследования.

Шаг 8. Выполняется ранжирование сформированного списка из 16 факторов с использованием метода пошагового уточнения ранжирования объектов [1-5].

2.Результаты апробации метода выделения искомого подмножества объектов из множества большой мощности. При проведении исследований, связанных с выделением факторов, влияющих на показатели уровня жизни населения субъектов РФ, оказалось, что количество социально-экономических показателей, которые теоретически могут влиять на уровень жизни населения, приближается к ста. Причем обнаружилось, что и статистически значимое влияние оказывают несколько десятков факторов [6]. Поэтому возникла необходимость *поиска* минимальной совокупности наиболее значимых, определяющих факторов для построения прогнозных и/или оптимизационных моделей путем выделения искомого подмножества объектов из множества достаточно большой мощности. Реализация этого метода позволила сформировать оригинальный, содержательно обоснованный состав независимых переменных и построить модели, обладающие хорошими прогнозными свойствами [7-10]. Причем все модели построены по данным Росстата до 2015 года включительно, т.к. в сборнике Росстата за 2018 год сведения о социально-экономических показателях развития субъектов РФ по 2016 году представлены не в полном объеме. Поэтому после того, как в 2019 году Росстатом издан сборник с данными за 2016 год, нами выполнена оценка прогнозных свойств ранее построенных моделей по нескольким показателям, в т.ч. по показателю «Среднемесячная номинальная начисленная заработная плата работников в субъектах РФ», т.к. по этому показателю в новом сборнике есть фактические данные и за 2016, и за 2017 годы (см. ниже гистограммы ошибки прогноза по субъектам РФ в 2016 и 2017 годах и прогноз на 2018 год). Результаты такой оценки качества построенных моделей оказались достаточно успешными (см. рисунки 1 и 2).

Так, средняя ошибка нашего прогноза по представленным в новом сборнике Росстата фактическим сведениям по всем 85 субъектам РФ составляет 2,1% за 2016 год и лишь 1,65% (!) за 2017 год. Но если исключить даже менее 5% явных выбросов, что вполне допустимо (см., например, [11]), то ошибки станут ещё меньше. А в ближайшем будущем предстоит очередная апробация метода путем выделения ограниченного подмножества объектов (показателей социально-экономического развития стран мира) из исходного множества достаточно большой мощности (содержащего более тысячи объектов-показателей), сформированного Всемирным банком.

Рис. 1 – Ошибка прогноза Среднемесячной номинальной начисленной заработной платы работников в субъектах РФ в 2016 году

Рис. 2. - Ошибка прогноза Среднемесячной номинальной начисленной заработной платы работников в субъектах РФ в 2017 году

Одновременно мы попытались спрогнозировать величину Среднемесячной номинальной начисленной заработной платы работников в субъектах РФ в 2018 году – см. рисунок 3 (Росстат опубликует фактические данные лишь в 2020 году).

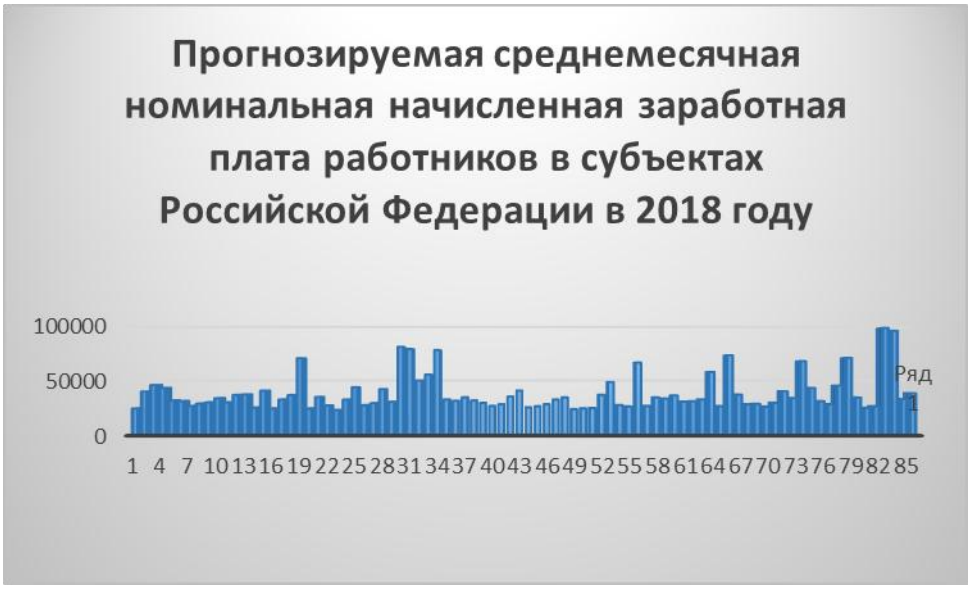

Рис. 3. – Прогнозируемая среднемесячная номинальная заработная плата работников в субъектах РФ в 2018 году (Росстат опубликует фактические данные в 2020 году).

ВЫВОДЫ. 1.Предложен и успешно апробирован оригинальный метод выделения искомого (весьма ограниченного) подмножества объектов из множества достаточно большой мощности исходных объектов.

2.Результаты апробации метода на открытых официальных статистических данных подтвердили перспективную полезность его применения в различных предметных областях.

3.Использование метода для оценки и прогнозирования различных показателей уровня жизни населения субъектов РФ позволило построить регрессионные модели *лучшего качества (при оценке по стандартным критериям статистической значимости - $R^2_{скорр}>0,9$; $F_{кр}$ >>100, в большинстве случаев у b–коэффициентов отношение $b_i/Б_{bi}$ >>2); *с лучшими прогнозными свойствами; *с использованием открытых официальных статистических данных.

Автор благодарен доктору экономических наук Игорю Сергеевичу Иванченко за подробный критический анализ публикаций в зарубежных изданиях по теме исследования, программистам Дмитрию Сергеевичу Сидоренко и Капитолине Николаевне Щербаковой за помощь в разработке и администрировании веб-приложения: URL: http://uroven-zhizni.ru.

Литература

1.Хубаев Г.Н. Об одном методе получения и формализации априорной информации при отборе значимых факторов// Сб. докладов итоговой науч. конф. Рост. ин-та народн. хоз-ва. Вып. 1. – Ростов-на-Дону, 1973. – С. 238-244.

2.Хубаев Г.Н. Математические методы и вычислительная техника в задачах упорядочения объектов и при отборе значимых факторов: Учеб. пособие / Под ред. А.Я. Боярского. – Ростов-на-Дону: РИНХ, 1975. – С.12-23.

3.Хубаев Г.Н. Математическое моделирование на предприятии: Учеб. пособие. – Ростов-на-Дону, 1973. – С. 84-89.

4.Хубаев Г.Н. Алгоритмы классификации лиц, принимающих решения, по уровню профессиональных знаний и творческим способностям // Наука и мир. – 2016. – № 5 (33). Ч.2. – С. 168-176.

5.Khubaev G. Expert review: method of intuitively agreed choice // 5th International Conference «Economy modernization: new challenges and innovative practice» (November 12, 2017, Sheffield, UK). – p. 65-80.

6.Хубаев Г.Н. Уровень бедности населения субъектов Российской Федерации: выявление, исследование и оценка статистической значимости определяющих факторов // РИСК: Ресурсы, информация, снабжение, конкуренция. – 2018. – № 3. – С. 72-75.

7.Хубаев Г.Н. Регрессионные модели для прогнозирования продолжительности жизни населения административно-территориальных образований: построение и оценка качества (Khubaev G. Regression models for forecasting life period of population of administrative-territorial education: construction and evaluation of quality) // Бюллетень науки и практики. – 2018. – Т. 4. № 9. – С. 206-217.

8.Сайт: «Субъекты РФ: анализ динамики социально-экономических показателей». URL: http://uroven-zhizni.ru.

9.Хубаев Г.Н. Качество жизни населения административно-территориальных образований: методика экспресс-анализа // Системный анализ в проектировании и управлении (SAEC-2018): Сборник научных трудов XXII Международной научно-практической конференции (г. Санкт-Петербург, СПбПУ им. Петра Великого, 22-24 мая 2018 г.). – СПб.: Изд-во Политехн. ун-та, 2018. – С. 139-146.

10.Хубаев Г.Н. Уровень жизни населения субъектов Российской Федерации: статистическое исследование // Статистика – язык цифровой цивилизации: сборник докладов Международной научно-практической конференции «II Открытый российский статистический конгресс» (г. Ростов-на-Дону, 4-6 декабря 2018 г.): в 2 т. – Т. 1. / Российская ассоциация статистиков; Федеральная служба государственной статистики РФ, РГЭУ (РИНХ), Ростовское региональное отделение ВЭО России. – Ростов н/Д, 2018. – С. 409-414.

11.Закс Л. (Lothar Sachs). Статистическое оценивание. / Пер. с нем. В.Н. Варыгина. – М.: «Статистика», 1976. – 598 с.

Literature

1. Hubayev G.N. About one method of receiving and formalization of prior information at selection of significant factors//Sb. reports total науч. конф. Growth. Inta народн. hoz-va. Issue 1. – Rostov-on-Don, 1973. – Page 238-244.

2. Hubayev G.N. Mathematical methods and computer facilities in problems of streamlining of objects and at selection of significant factors: Studies. a grant / Under the editorship of A.Ya. Boyarsky. – Rostov-on-Don: RINH, 1975. – Page 12-23.

3. Hubayev G.N. Mathematical modeling at the enterprise: Studies. grant. – Rostov-on-Don, 1973. – Page 84-89.

4. Hubayev G.N. Algorithms of classification of the persons making decisions on the level of professional knowledge and creative abilities//Science and the world. – 2016. – No. 5 (33). Ch.2. – Page 168-176.

5.Khubaev G. Expert review: method of intuitively agreed choice // 5th International Conference «Economy modernization: new challenges and innovative practice» (November 12, 2017, Sheffield, UK). – p. 65-80.

6. Hubayev G.N. Level of poverty of the population of territorial subjects of the Russian Federation: identification, a research and assessment of the statistical importance of the defining factors//RISK: Resources, information, supply, competition. – 2018. – No. 3. – Page 72-75.

7. Hubayev G.N. Regression models for forecasting of life expectancy of the population of administrative-territorial educations: construction and assessment of quality (Khubaev G. Regression models for forecasting life period of population of administrative-territorial education: construction and evaluation of quality)//Bulletin of science and practice. – 2018. – T. 4. No. 9. – Page 206-217.

8. Website: "Territorial subjects of the Russian Federation: analysis of dynamics of socio-economic indexes". URL: http://uroven-zhizni.ru.

9. Hubayev G.N. Quality of life of the population of administrative-territorial educations: an express analysis technique//the System analysis in design and management (SAEC-2018): The collection of scientific works of the XXII International scientific and practical conference (St. Petersburg, СПбПУ of Peter the Great, on May 22-24, 2018). – SPb.: Politekhn publishing house. un-that, 2018. – Page 139-146.

10. Hubayev G.N. Standard of living of the population of territorial subjects of the Russian Federation: a statistical research//Statistics – language of a digital civilization: collection of reports of the International scientific and practical conference "the II Open Russian statistical congress" (Rostov-on-Don, on December 4-6, 2018): in 2 t. – T. 1. / Russian association of statisticians; Federal State Statistics Service of the Russian Federation, RGEU (RINH), Rostov regional office of VEO of Russia. – Rostov N / Д, 2018. – Page 409-414.

11. Legislative Assembly of L. (Lothar Sachs). Statistical estimation. / The lane with it. V.N. Varygina. – M.: "Statistics", 1976. – 598 pages.