РАЗДЕЛ III
МЕТОДИКИ АНАЛИЗА И ПРОГНОЗИРОВАНИЯ РИСКОВ СОЦИАЛЬНО-ПОЛИТИЧЕСКОЙ ДЕСТАБИЛИЗАЦИИ

ОЦЕНКА ПОСТКРИЗИСНОГО РАЗВИТИЯ СОЦИАЛЬНО-ПОЛИТИЧЕСКИХ СИСТЕМ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ*

Сергей Георгиевич Шульгин
Российская академия народного хозяйства и государственной службы при Президенте РФ;
Национальный исследовательский университет «Высшая школа экономики»

В статье предложен алгоритм для оценки посткризисного развития социально-политических систем. Для каждого индекса нестабильности на выбранном временном интервале мы оцениваем среднее значение индекса и среднеквадратическое отклонение. Моментом кризиса считаем ситуацию превышения значения индекса выше порогового уровня (среднее значение плюс среднеквадратическое отклонение), посткризисный анализ включает интервал до момента, когда значение индекса опускается ниже среднего значения. С использованием методов машинного обучения мы проанализировали динамику посткризисного развития социально-политических систем. В статье мы определяем типы посткризисного развития и предлагаем алгоритм для анализа факторов, которые оказывают влияние на посткризисную динамику. Анализ мы проводили на основе базы данных социально-политических факторов, включающей более 400 независимых переменных, для 25 тыс. наблюдений в формате «страна – год». Основной источник сведений по социально-политической нестабильности – данные The Cross-National Time Series (CNTS). В качестве источников данных мы использовали: World Bank, Worldwide Governance Indicators, Polity IV, Maddison Database, CNTS, United Nation Population Division, World Value Survey и др.

* Исследование выполнено при поддержке Российского научного фонда (проект № 18-18-00254).

Введение

В серии предыдущих работ (Шульгин 2018; 2019) мы использовали методы машинного обучения для отбора и анализа факторов, влияющих на социально-политическую динамику. В данной работе мы используем аналогичные методы для решения другой задачи – классификации типов посткризисной динамики и также анализируем набор факторов, влияющих на возможную посткризисную динамику.

Изучение посткризисной динамики может быть основано на обширном экономико-историческом материале (см., например: Goldstone 2016; Epstein *et al.* 2006; May, Стародубровская 2001; Mau 2017). Авторы предлагают как различные структурныt теории социально-политических потрясений, так и их систематический анализ и классификацию.

Традиционный подход к анализу факторов социально-политической нестабильности предполагает отбор факторов, которые теоретически могут влиять на варианты посткризисной динамики. Группами факторов, которые могут определять варианты посткризисной динамики, являются: социально-экономические характеристики общества; характеристики политического устройства; тип государственного управления; социально-демографические факторы; история политического режима, его устойчивость и т. п.

Подробнее классификация факторов была описана в статье (Шульгин 2019). Мы приводили обзор отдельных подходов к анализу нестабильности, который представлен в работе 2017 г. (Коротаев и др. 2017). Анализ отдельных факторов представлен, например, в работах (Esty *et al* 1998; Цирель 2012; 2015; Коротаев, Зинькина 2012; Малков и др. 2013; Przeworski *et al.* 2000) и множестве других.

Анализ социально-политических процессов возможен и с использованием микроданных (отдельных событий, высокой географической детализации, поведения отдельных людей и т. п.), для работы с которыми используются методы машинного обучения. Обзор методов представлен в работе 2017 г. (Donnay 2017), а примерами исследований являются работы (Connelly *et al.* 2016; Donnay *et al.* 2016; Coppock *et al.* 2016) и др.

В этой статье мы применяем методы машинного обучения для эмпирического анализа процессов посткризисной динамики. Мы используем данные на страновом уровне с разбивкой по годам. Та-

кая градация позволяет объединять различные источники данных и использовать широкие массивы независимых переменных и проанализировать их влияния на посткризисную динамику.

Модель

Для заданного набора данных D определены n точек данных, при этом каждая точка данных – это набор из объясняемой (зависимой) переменной y_i и множества из m независимых факторов X_i:

$D = \{(y_i, X_i)\}$ ($|D| = n$, $X_i \in \mathbb{R}^m$, $y_i \in \mathbb{R}$) (

Где $\mathbb{R}$ – стандартное обозначение для множества действительных чисел.

В такой формулировке наша задача – среди всего множества независимых факторов X выделить такое его подмножество, то есть отдельные его факторы, которые оказываются наиболее важными для объяснения y.

В данной работе мы используем метод, при котором мы пытаемся найти оценку зависимой переменной y_i в форме K аддитивных функций:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i). \quad ($$

$f_k(X_i)$ – функция, которая принадлежит к подмножеству классификационных и регрессионных деревьев (CART – *Classification and Regression Tree*).

Класс функций, которые определяются как:

$CART = \{f(X) = w_{q(X)}\}$ ($q: \mathbb{R}^m \to T$; $w \in \mathbb{R}^T$), (

где q(X) описывает дерево, вершинами которого являются правила относительно значений X. Функция q(X) ставит в соответствие определенной точке данных X_i определенный лист (конечную вершину) (T). Листья в CART описывают результат классификации, которым присвоены веса w. Аппроксимирующая функция $f_k(X)$ определяется структурой дерева q(X) и весами листьев w.

Процесс обучения (тренировки) модели сводится к минимизации функционала L, в которой суммируется ошибка между оцененными ($\hat{y}_i$) и реальными значениями (y_i) зависимой переменной, а также учитывается сложность (размерность) CART-функции. Вторая часть функционала L – это элемент так называемой регуляризации, подход, с помощью которого мы контролируем сложность CART-функции и пытаемся найти самую простую структуру из возможных CART-функций.

Для минимизации функционала L используется последовательный (итеративный) процесс, где на каждой итерации оценивается градиент в направлении минимизации L (подробнее описание функционала и алгоритма оптимизации см.: Chen, Guestrin 2016).

Использование моделей градиентного бустинга (GBM) не требует нормализации данных для корректной работы и хорошо работает без предварительной обработки входных данных. В работе мы использовали GBM для всех оценок, которые производили с помощью библиотеки XGBoost (*Ibid.*).

Данный метод успешно применяется для широкого класса задач, связанного с отбором наиболее важных переменных в задачах с высокой размерностью. Например, в отборе оптимальных характеристик соискателей для предсказания для них наиболее интересных и релевантных вакансий (Volkovs *et al.* 2017), или предсказания о том, какие наиболее значимые аффилиации авторов влияют на факт, что их статьи принимаются на основные авторитетные научные конференции в области машинного обучения, больших данных и т. п. (Sandulescu, Chiru 2016), или анализе физических данных CERN, полученных на Большом адронном коллайдере, в попытках найти факторы, влияющие на вероятность наблюдения редкого физического явления – распада тау-лептона на три мюона ($\tau \rightarrow 3\mu$) (Mironov, Guschin 2015) и во многих других приложениях.

Данные

В качестве исходных данных о нестабильности мы используем данные *Cross National Time Series* (CNTS), *Global Terrorism Database* (GTB) и базы данных государственных переворотов.

База данных *The Cross National Time Series* (CNTS) – это результат работы по сбору и систематизации данных, начатой Артуром Банксом (Banks, Wilson 2020) в 1968 г. в Университете штата Нью-Йорк в Бингемтоне, обобщающей архив данных The Statesman's Yearbook, публикуемых с 1864 г. В базе содержатся данные по более чем 200 странам, годовые значения переменных начиная с 1815 г. В базе данных исключены периоды двух мировых войн 1914–1918 и 1940–1945 гг.

В данной работе мы используем в качестве зависимых переменных данные, описывающие различные аспекты внутренних конфликтов (domestic). Эти данные получены из анализа страновых событий по 8 различным подкатегориям:

- Политические убийства (*Assassinations*, domestic1).
- Политические забастовки (*General Strikes*, domestic2).
- Партизанские действия (*Guerrilla Warfare*, domestic3).
- Правительственные кризисы (*Government Crises*, domestic4).
- Политические репрессии (*Purges*, domestic5).
- Массовые беспорядки (*Riots*, domestic6).
- Перевороты и попытки переворотов (Revolutions, domestic7).
- Антиправительственные демонстрации (Anti-Government Demonstrations, domestic8).

К «Политическим убийствам» (*Assassinations*, domestic1) относятся любые политически мотивированные убийства или покушения на убийства высших правительственных чиновников или политиков.

К «Политическим забастовкам» (*General Strikes*, domestic2) относятся забастовки, в которых участвовало 1000 или более работников, более одного работодателя и при этом звучали требования, направленные против национальной политики, правительства или органов власти.

К «Партизанским действиям» (*Guerrilla Warfare*, domestic3) относится любая вооруженная деятельность, диверсии или взрывы, совершаемые независимыми группами граждан или нерегулярными вооруженными силами, которые направлены на свержение нынешнего режима.

К «Правительственным кризисам» (*Government Crises*, domestic4) относятся любые ситуации, которые грозят привести к падению текущего режима – за исключением вооруженных переворотов, напрямую направленных на это.

К «Политическим репрессиям» (*Purges*, domestic5) относятся любые систематические устранения политической оппозиции (лишения свободы или убийства) среди действующих членов режима или политической оппозиции.

К «Массовым беспорядкам» (*Riots*, domestic6) относятся любые демонстрации или столкновения, связанные с использованием насилия, в которых принимали участие более 100 граждан.

К «Переворотам и попыткам переворотов» (*Revolutions*, domestic7) относятся любые незаконные или связанные с принуждением изменения в правящей элите, а также любые попытки таких изменений. Переменная «Перевороты и попытки переворотов» также учитывает все удачные и неудачные вооруженные восстания, це-

лью которых является получение независимости от центрального правительства.

К «Антиправительственным демонстрациям» (*Anti-Government Demonstrations*, domestic8) относятся любые мирные публичные собрания, в которых принимает участие 100 и более человек, а основной целью проведения является выражение несогласия с политикой правительства или власти за исключением демонстраций с выраженным направлением против иностранных государств.

Все перечисленные 8 подкатегорий используются при построении общего индекса социально-политической стабилизации (domestic9). Для этого составители базы данных *CNTS* присвоили каждой подкатегории определенный вес (см. Табл. 1).

Табл. 1. Веса подкатегорий, используемых при построении индекса социально-политической стабилизации

Подкатегория	Название переменной	Вес в индексе социально-политической стабилизации (domestic9)
Политические убийства (*Assassinations*)	cnts_domestic1	25
Политические забастовки (*General Strikes*)	cnts_domestic2	20
Партизанские действия (*Guerrilla Warfare*)	cnts_domestic3	100
Правительственные кризисы (*Government Crises*)	cnts_domestic4	20
Политические репрессии (*Purges*)	cnts_domestic5	20
Массовые беспорядки (*Riots*)	cnts_domestic6	25
Перевороты и попытки переворотов (*Revolutions*)	cnts_domestic7	150
Антиправительственные демонстрации (*Anti-Government Demonstrations*)	cnts_domestic8	10

Индекс социально-политической стабилизации (*Weighted Conflict Measure*, domestic9) рассчитывается по формуле (4):

$$domestic9 = \frac{\sum_{i=1}^{8} w_i \, \text{cnts\_}domestic_i}{8} * 100,$$

где w_i – веса, приведенные в последнем столбце Табл. 1.

Кроме показателя domestic9 для анализа мы построили переменную domestic9 с лагом (cnts_domestic9_prev), которая показывает общее значение страновой нестабильности в предыдущем году. Также мы построили упреждающую переменную (cnts_domestic9_next) для оценки общего уровня нестабильности в будущем году.

Помимо данных *CNTS*, в качестве объясняемой переменной мы используем два индикатора из *Global Terrorism Database* (START 2020). Мы используем переменные:

n_terror_attack – количество террористических атак,

Nkill – количество убитых.

База содержит данные с 1970 г. (в анализируемой версии по 2015 г. включительно)

Из базы данных государственных переворотов (Marshall 2016) для независимых переменных мы взяли для анализа переменную:

coup_detat_failed_coup_detat – государственные перевороты и попытки переворотов (аналог переменной cnts_domestic8).

База данных государственных переворотов охватывает временной период с 1960 по 2016 г.

Методология

Понятие посткризисной динамики мы определяем через понятие социально-политического кризиса. Социально-политический кризис в определенной стране мы определяем как момент времени, в который количественные показатели социально-политической нестабильности выходят за границы порогового уровня, описывающего страновую динамику социально-политической нестабильности. Для этого по временным рядам мы для каждой страны оцениваем среднее значение и среднеквадратическое отклонение.

Мы выделяем в отдельную группу зависимых переменных набор факторов, которые отражают те или иные формы социально-политической нестабильности. Для поиска наиболее важных факторов мы тренируем (оцениваем) множество моделей, в каждой из которых лишь одна из зависимых переменных используется в качестве целевой объясняемой переменной.

Всего в качестве зависимых (объясняемых, целевых) для данного анализа было отобрано 12 переменных. Точки данных, в которых значение зависимой переменной оказывалось больше порогового уровня (среднего значения плюс одно среднеквадратическое отклонение), были классифицированы как начало кризиса. На Рис. 1 приведены несколько страновых рядов данных.

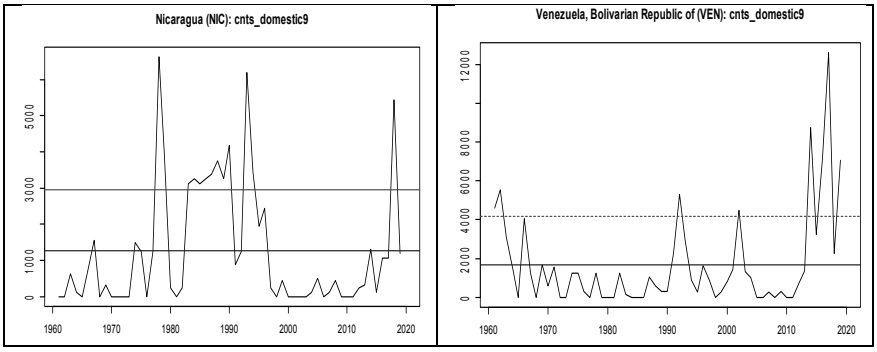

Рис. 1. Страновые ряды данных

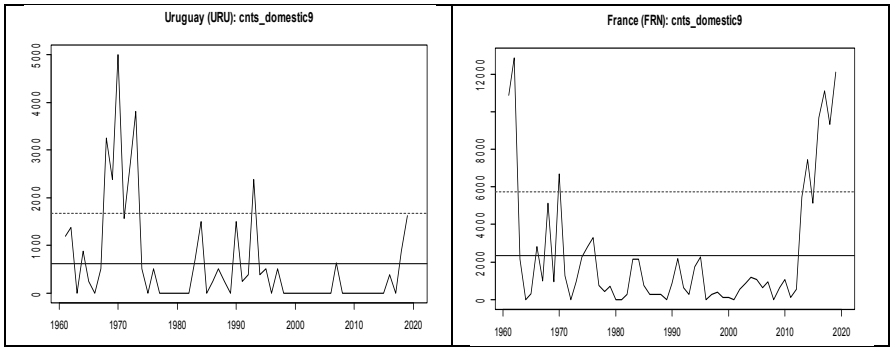

Рис. 2. Примеры временных рядов политической нестабильности для отдельных стран с использованием переменной domestic_9

В Табл. 2 приведены статистика по 12 зависимым переменным, число случаев социально-политических кризисов и длительность кризисной для анализируемых переменных:

Табл. 2. Статистика анализируемых кризисов и их продолжительности

Переменная	Число наблюдений (страна-год)	Число кризисных лет	Число кризисов	Число кризисов длительностью (лет)				
				1	2	3	4	5 и больше
cnts_domestic1	10183	578	368	261	66	18	10	13
cnts_domestic2	10183	556	365	259	66	21	7	12
cnts_domestic3	10183	639	286	151	60	24	18	33
cnts_domestic4	10183	917	630	473	95	31	16	15
cnts_domestic5	10183	538	384	290	62	21	3	8
cnts_domestic6	10183	970	544	347	93	37	38	29
cnts_domestic7	10183	1018	476	296	82	40	17	41
cnts_domestic8	10183	1042	521	302	93	45	37	44
cnts_domestic9	10183	1357	568	281	123	54	28	82
coup_detat_failed_coup_detat	12392	404	310	250	34	20	4	2
n_terror_attack	9612	1145	453	216	90	44	34	69
Nkill	9612	811	358	198	64	32	17	47

Мы ограничились периодом с 1961 по 2019 г. Столбец «число наблюдений» показывает количество непустых точек данных (страна – лет), столбец «число кризисных лет» показывает общее число кризисных лет (точек данных), в которых в стране начался кризис (первысил пороговое значение начала кризиса) и еще не закончился (не опустился ниже порогового значения окончания кризиса). Столбец «число кризисов» показывает общее число кризисов, включая текущие, которые еще не закончились. Мы оценили для всех баз данных, какое количество кризисов длилось 1 год, 2 года, 3 года, 4 года, 5 и более лет.

Разработка модели анализа посткризисного развития

Используя размеченные данные, в момент наступления кризиса мы можем обучить модель, используя полный набор независимых переменных, оценивать тип предстоящего кризиса и возможную посткризисную динамику. Для оценки (тренировки) модели градиентного бустинга необходимо выбрать набор параметров, определяющих работу алгоритма. Одна из главных проблем, которые необходимо решить при оценке, – это проблема переобучения мо-

дели (over-fitting). Переобучение выражается в том, что при большом количестве данных и степеней свободы модель может очень точно описать существующие закономерности на обучающей выборке (training set), однако полученные закономерности могут оказаться неприменимы за пределами обучающей выборки.

Подробнее методология и параметры используемых моделей описаны в работе (Шульгин 2019).

Заключение и обсуждение результатов

Проведена типология посткризисной динамики социально-политических систем. Проведен страновой анализ временных рядов индексов социально-политической нестабильности и предложен набор пороговых уровней, характеризующих начало кризиса и момент окончания поскризисного периода. В качестве порогового значения для момента начала выбрана сумма среднего значения и среднеквадратического отклонения для анализируемого индекса социально-политической нестабильности на временном интервале (1961–2019 гг.). Разработан алгоритм для анализа факторов, которые оказываются наиболее важными для понимания вариантов посткризисной социально-политической динамики.

Библиография

Коротаев А. В., Зинькина Ю. В. 2012. Структурно-демографические факторы «арабской весны». *Протестные движения в арабских странах. Предпосылки, особенности, перспективы* / Ред. И. В. Следзевский, А. Д. Саватеев. М.: ЛИБРОКОМ/URSS. С. 28–40.

Коротаев А. В., Шульгин С. Г., Зинькина Ю. В. 2017. *Анализ страновых рисков с использованием демографических и социально-экономических данных*. М.: РАНХиГС. URL: http://dx.doi.org/10.2139/ssrn.2944064.

Малков С. Ю., Коротаев А. В., Исаев Л. М., Кузьминова Е. В. 2013. О методике оценки текущего состояния и прогноза социальной нестабильности: опыт количественного анализа событий Арабской весны. *Полис. Политические исследования* 4: 137–162.

Мау В. А., Стародубровская И. В. 2001. *Великие революции от Кромвеля до Путина*. М.: Дело.

Цирель С. В. 2012. Условия возникновения революционных ситуаций в арабских странах. Арабская весна 2011 года. *Системный мониторинг глобальных и региональных рисков:* ежегодник. Т. 3. *Арабская весна*

2011 года / Отв. ред. А. В. Коротаев, Ю. В. Зинькина, А. С. Ходунов. М.: ЛИБРОКОМ/URSS. С. 162–173.

Цирель С. В. 2015. К истокам украинских революционных событий 2013–2014 гг. *Системный мониторинг глобальных и региональных рисков:* ежегодник. Т. 6. *Украинский разлом* / Отв. ред. Л. Е. Гринин, А. В. Коротаев, Л. М. Исаев, А. Р. Шишкина. Волгоград: Учитель. С. 57–83.

Шульгин С. Г. 2018. Отбор переменных для анализа и прогнозирования нестабильности с помощью моделей градиентного бустинга. *Системный мониторинг глобальных и региональных рисков:* ежегодник. Т. 9. *Социально-политическая и экономическая дестабилизация: анализ страновых и региональных ситуаций в мир-системном аспекте* / Отв. ред. Л. Е. Гринин, А. В. Коротаев, К. В. Мещерина. Волгоград: Учитель. С. 115–153.

Шульгин С. Г. 2019. Анализ факторов социально-политической нестабильности в странах Афразийской макрозоны с помощью моделей машинного обучения. *Системный мониторинг глобальных и региональных рисков:* ежегодник. Т. 10 / Отв. ред. Л. Е. Гринин, А. В. Коротаев, К. В. Мещерина. Волгоград: Учитель. С. 188–226.

Banks A. S., Wilson K.A. 2020. *Cross-National Time-Series Data Archive*. Jerusalem, Israel: Databanks International. URL: https://www.cntsdata.com/

Chen T., Guestrin C. 2016. Xgboost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACMSIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM. Pp. 785–794.

Chen T., He T., Benesty M., Khotilovich V.,Tang Y., Cho H., Chen K., Mitchell R., Cano I., Zhou T., Li M., Xie J., Lin M., Geng Y., and Li Y. 2018. *xgboost: Extreme Gradient Boosting. R Package Version 0.71.2*. URL: https://CRAN.R-project.org/package=xgboost.

Connelly R., Playford C. J., Gayle V. and Dibben C. 2016. The Role of Administrative Data in the Big Data Revolution in Social Science Research. *Social Science Research* 59: 1–12. DOI: 10.1016/j.ssresearch.2016.04.015.

Coppock A., Guess A., Ternovski J. 2016. When Treatments are Tweets: A Network Mobilization Experiment over Twitter. *Political Behavior* 38(1): 105–128. DOI : 10.1007/s11109-015-9308-6.

Donnay K. 2017. Big Data for Monitoring Political Instability. *International Development Policy. Revue internationale de politique de développement* 8 (8.1).

Donnay K., Dunford E., McGrath E. C., Backer D., Cunningham D. E. 2016. *MELTT: Matching Event Data by Location, Time and Type*. Paper

presented at the Annual Conference of the Midwest Political Science Association, Chicago.

Epstein D. L., Bates R., Goldstone J., Kristensen I., O'Halloran S. 2006. Democratic Transitions. *American Journal of Political Science* 50(3): 551–569.

Esty, D., Goldstone J. A., Gurr T. R., Harff B., Levy M., Dabelko G. D., Surko P., Unger A. N. 1998. State Failure Task Force Report: Phase II Findings. McLean, VA: Sci. Appl. Int. Corp.

Goldstone J. A. 2016. *Revolution and Rebellion in the Early Modern World: Population Change and State Breakdown in England, France, Turkey, and China, 1600–1850.* Routledge.

Kuhn M. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* 28(5): 12–6. DOI: http://dx.doi.org/10.18637/jss.v028.i05.

Marshall M. G. 2016. *Coup D'État Events, 1946-2015 Codebook* / Ed. by M. G. Marshall, D. R. Marshall. Center for Systemic Peace.

Mau V. 2017. *Russia's Economy in an Epoch of Turbulence: Crises and Lessons*. Routledge.

Mironov V., A. Guschin. 2015. *1st place of the CERN LHCb Experiment Flavour of Physics Competition.* URL: http://blog.kaggle.com/2015/11/30/flavour-of-physics-technical-write-up-1st-place-go-polar-bears/.

Przeworski A., Alvarez M. E., Cheibub J. A., Limongi F. 2000. Democracy and Development. *Political Institutions and Well-Being in the World, 1950–1990* / Ed. by A. Przeworski. New York: Cambridge University Press.

Sandulescu V., Chiru M. 2016. *Predicting the Future Relevance of Research Institutions – The Winning Solution of the KDD Cup 2016*. URL: arXiv preprint arXiv:1609.02728.

START [National Consortium for the Study of Terrorism and Responses to Terrorism]. 2020. *Global Terrorism Database*. College Park, MD: University of Maryland. URL: https://www.start. umd.edu/gtd.

Volkovs M., Yu G. W., Poutanen T. 2017. Content-based Neighbor Models for Cold Start in Recommender Systems. *Proceedings of the Recommender Systems Challenge.* ACM. P. 7. DOI: https://doi.org/10.1145/3124791.3124792.