DOI: 10.14515/monitoring.2022.2.1921

А. А. Воробьев, Е. Д. Ноздрачева

МЕТОДИКА ПРОГНОЗИРОВАНИЯ РЕЗУЛЬТАТОВ ВЫБОРОВ С ИСПОЛЬЗОВАНИЕМ КОМБИНИРОВАНИЯ МЕТОДОВ АНАЛИТИЧЕСКОГО И ТРЕНДОВОГО МОДЕЛИРОВАНИЯ

Правильная ссылка на статью:
Воробьев А. А., Ноздрачева Е. Д. Методика прогнозирования результатов выборов с использованием комбинирования методов аналитического и трендового моделирования // Мониторинг общественного мнения: экономические и социальные перемены. 2022. № 2. С. 24—41. https://doi.org/10.14515/monitoring.2022.2.1921.
For citation:
Vorobyev A. A., Nozdracheva E. D. (2022) Methodology for Forecasting Election Results Using Combination of Analytical and Trend Modeling Methods. *Monitoring of Public Opinion: Economic and Social Changes.* No. 2. P. 24–41. https://doi.org/10.14515/monitoring.2022.2.1921. (In Russ.)

DOI: 10.14515/monitoring.2022.2.1921

МЕТОДИКА ПРОГНОЗИРОВАНИЯ РЕЗУЛЬТАТОВ ВЫБОРОВ С ИСПОЛЬЗОВАНИЕМ КОМБИНИРОВАНИЯ МЕТОДОВ АНАЛИТИЧЕСКОГО И ТРЕНДОВОГО МОДЕЛИРОВАНИЯ

ВОРОБЬЕВ Андрей Анатольевич — кандидат технических наук, доцент, сотрудник, Академия ФСО России, Орел, Россия
E-MAIL: awa@mail.ru
https://orcid.org/0000-0002-1566-4139

НОЗДРАЧЕВА Екатерина Дмитриевна — сотрудник , Академия ФСО России, Орел, Россия
E-MAIL: katerina_belonozhko@mail.ru
https://orcid.org/0000-0001-6123-4060

Аннотация. Работа посвящена исследованию возможностей прогнозирования рейтинга политических партий на коротких временных рядах с помощью данных социологических опросов ВЦИОМ по выборной тематике. Рассмотрены три основные группы методов: трендовое моделирование, экспертные оценки, аналитическое моделирование.

В результате теоретического анализа и проведенных экспериментов выявлена низкая точность прогнозов, полученных только с помощью методов, относящихся к группе трендового моделирования. Поэтому предложена методика, позволяющая повысить точность прогнозирования за счет корректировки полученных в результате логарифмической аппроксимации прогнозов. С помощью иерархической кластеризации находится вектор значений, состоящий из коэффициентов уравнений аппроксимации и описывающий наиболее

METHODOLOGY FOR FORECASTING ELECTION RESULTS USING COMBINATION OF ANALYTICAL AND TREND MODELING METHODS

Andrey A. VOROBYEV[1] *— Cand. Sci. (Tech.), Associate Professor, employee*
E-MAIL: awa@mail.ru
https://orcid.org/0000-0002-1566-4139

Ekaterina D. NOZDRACHEVA[1] *— employee*
E-MAIL: katerina_belonozhko@mail.ru
https://orcid.org/0000-0001-6123-4060

[1] Russian Federation Security Guard Service Federal Academy, Oryol, Russia

Abstract. The study is devoted to forecasting political party ratings using short time series based on the VCIOM electoral polls data. The authors examine three groups of methods: trend modeling, expert assessments, and analytical modeling.

A review of the theoretical background and the results of empirical experiments shows low accuracy of forecasts produced solely by the trend modeling methods. To improve their accuracy, the authors propose a method of adjusting forecasts by logarithmic approximation. This method bases on hierarchical clustering that uses a vector of the coefficients of the approximated equations and describes the most similar electoral situation in the past. Then the reminder calculated as the difference between the sum of the predicted values and 100% is proportionally redistributed among the participants of the election campaign. To forecast political party ratings where the

DOI: 10.14515/monitoring.2022.2.1921

«похожую» электоральную ситуацию в прошлом. Затем «остаток», вычисленный как разница суммы прогнозных значений и 100 %, пропорционально перераспределяется в соответствии с участниками избирательной кампании. Для прогнозирования рейтинга партий, где значения временного ряда не превышают 5 %, предпочтительнее применять метод среднего темпа роста.

Ключевые слова: опросы общественного мнения, короткий временной ряд, интерполяция, трендовое моделирование, аппроксимация, иерархический кластерный анализ

time series values do not exceed 5%, it is preferable to use the average growth rate method.

Keywords: opinion polls, short time series, interpolation, trend modeling, approximation, hierarchical cluster analysis

Введение

Для изучения поведения электората в ходе избирательной кампании, которая в соответствии с действующим законодательством проводится в течение трех месяцев до дня голосования, а также для реализации функции контроля проведения выборов с целью минимизации возможности намеренного искажения результатов проводятся социологические опросы, на основе которых строится прогноз, позволяющий общественности ознакомиться с предварительными результатами выборов.

Один из актуальных методов сбора социологических данных — интервьюирование по репрезентативной выборке, не только учитывающей половозрастную структуру населения, но и охватывающей представителей всех социальных групп и населенных пунктов. Однако для достижения целевого показателя статистической ошибки выборки необходимы большие затраты человеческих, временных и материальных ресурсов. Поэтому такие социологические исследования проводятся, как правило, в течение месяца. Следовательно, за период избирательной кампании может собираться статистическая информация по результатам 3—4 последовательных социологических опросов, которые можно представить в виде короткого временного ряда.

Анализ методов прогнозирования [Горшков, 2011] на коротких временных рядах показывает низкую точность прогнозов по выборной тематике. Это связывают не только с ограниченностью длины ряда, но и использованием различных моделей прогноза, дающих разную степень достоверности тенденций в зависимости от явки избирателей. Кроме того, в результате анализа [Клисторин, 2011] было выявлено, что на точность прогнозов существенно влияют «неопределившиеся» респонденты. При большом количестве пропусков в данных, возникших в результате выбора варианта «затрудняюсь ответить», в оценках электоральных предпочтений могут возникать смещения.

DOI: 10.14515/monitoring.2022.2.1921

В работе предлагается решить проблему низкой точности прогнозов результатов выборов, построенных на коротких временных рядах, путем выдвижения и проверки гипотез, связанных с экспериментальным исследованием предлагаемых О. Капитановой трех групп методов: трендовое моделирование, экспертные оценки и аналитическое моделирование [Капитанова, 2016]. Основным результатом анализа является методическое обеспечение, позволяющее повысить точность прогнозных оценок за счет совместного использования методов, относящихся к группе трендового и аналитического моделирования.

Особенности используемых в работе социологических данных

В качестве исходных данных были использованы результаты социологических исследований Всероссийского центра изучения общественного мнения, проведенных в 2016 г. в ходе предвыборной кампании в Государственной думу Федерального Собрания Российской Федерации VII созыва по партийным спискам[1].

Известно, что на этапе предварительной подготовки исходных данных возможна их корректировка, основанная на моделях прогнозирования, учитывающих или не учитывающих наличие пропусков (неопределившихся респондентов) в данных [Горшков, 2011]. В работе использовались результаты экспериментов [Жучкова, Ротмистров, 2018], в которых данные с пропусками исключались из дальнейшего анализа, если количество пропущенных значений не превышало 5% от общего количества опрошенных. В случае превышения порога в 5% для перераспределения неопределившихся респондентов использовался метод множественной импутации, обоснованный в работе [Фабрикант, 2015] и реализуемый в виде методики, подробно рассмотренной в [Воробьев и др., 2020]. В состав методики входят следующие основные процедуры:

— выбор независимых переменных, коррелируемых с зависимой переменной («Партия»), с целью приведения неслучайных пропусков (NMAR) к случайным (MAR) с использованием подхода [Enders, 2010] и экспериментально проверенного в работах [Зангиева, 2011; Zhuchkova, Rotmistrov, 2022];

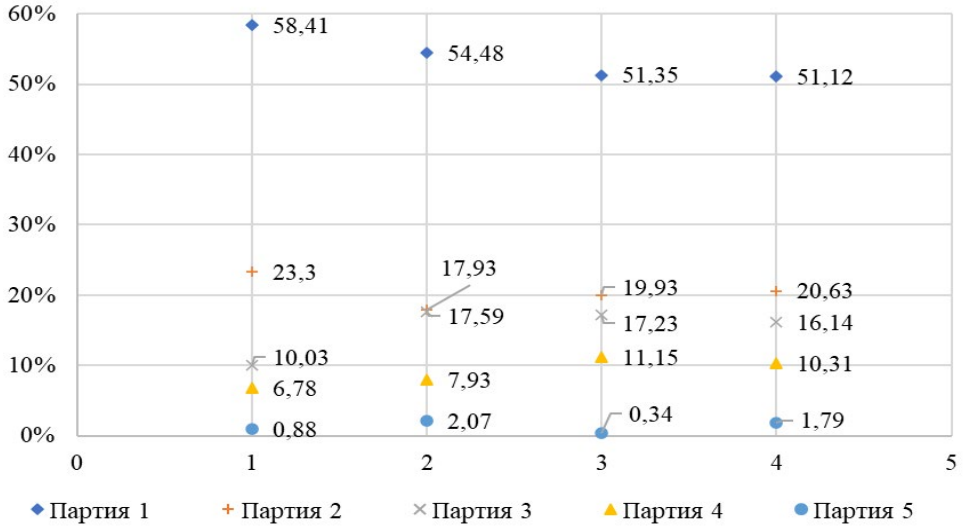
— расчет импутированных значений вместо пропусков с помощью процедуры «множественная импутация» (метод импутации «автоматический» обоснован в [Ворбьев, Воронецкий, 2019]), реализованной в программе IBM SPSS Statistics версии 23, где в качестве предикторов использовались отобранные в предыдущей процедуре независимые переменные.

На следующем этапе для создания временного ряда по каждой политической партии проводился частотный анализ результатов, полученных в рамках социологических опросов, реализовывавшихся в течение трех месяцев до выборов (по одному измерению ежемесячно) и непосредственно перед выборами (одно измерение). В связи с различным числом опрошенных респондентов в регионах все частоты переводились в процентные значения. На рисунке 1 представлено частотное распределение предпочтений электората (в процентах) в соответствии с месяцами опросов для одной из выборок.

[1] См. База социологических данных ВЦИОМ. URL: https://bd.wciom.ru/ (дата обращения: 15.02.2021).

DOI: 10.14515/monitoring.2022.2.1921

Рис. 1. Частотное распределение предпочтений электората в соответствии с месяцами опросов, в %



Однако в результате анализа [Домбровский, 2016] было определено, что минимальная длина короткого временного ряда для расчета одного прогнозного значения с помощью существующих методов прогнозирования должна быть не меньше 6. Следовательно, чтобы прогнозирование на полученном временном ряду было возможно, необходимо увеличить количество точек. В работе [Знаменский, 2018] было выявлено, что наиболее точной из существующих методов восстановления пропущенных значений во временном ряду является формула кусочно-квадратичной интерполяции:

$$F(x) = a_0 + a_1 x + a_2 x^2, \text{ при } x_{i-1} < x < x_{i+1}, \quad (1)$$

где коэффициенты a_0, a_1 и a_2 на каждом интервале $[x_{i-1}; x_{i+1}]$, x_i — частоты проголосовавших за каждого кандидата в i-й соцопрос, определяются решением системы уравнений для условия прохождения параболы через три точки:

$$\begin{cases} f_{i-1} = a_0 + a_1 x_{i-1} - a_2 x_{i-1}^2 \\ f_i = a_0 + a_1 x_i + a_2 x_i^2 \\ f_{i+1} = a_0 + a_1 x_{i+1} + a_2 x_{i+1}^2 \end{cases}, \quad (2)$$

Из системы уравнений находятся коэффициенты:

$$a_0 = f(x_{i-1}) - a_1 x_{i-1} - a_2 x_{i-1}^2, \quad (3)$$

$$a_1 = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} - a_2 (x_i + x_{i-1}), \quad (4)$$

DOI: 10.14515/monitoring.2022.2.1921

$$a_2 = \frac{f(x_{i+1}) - f(x_{i-1})}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)} - \frac{f(x_i) - f(x_{i-1})}{(x_i - x_{i-1})(x_{i+1} - x_i)}, \quad (5)$$

В конечное уравнение подставляются необходимые значения оси абсцисс и рассчитываются соответствующие им значения оси ординат.

Так был сформирован новый временной ряд (см. табл. 1), где в столбцах в названии с целыми числами представлены результаты непосредственно проведенных социологических опросов, а в столбцах с дробными значениями — интерполированные.

Таблица 1. ***Временной ряд с интерполированными значениями, построенный на одной из обучающих выборок***

	1	**1,5**	**2**	**2,5**	**3**	**3,5**	**4**
Партия 1	58,41	56,44	54,48	52,92	51,35	51,24	51,12
Партия 2	23,30	20,62	17,93	18,93	19,93	20,28	20,63
Партия 3	10,03	13,81	17,59	17,41	17,23	16,69	16,14
Партия 4	6,78	7,36	7,93	9,54	11,15	10,73	10,31
Партия 5	0,88	1,48	2,07	1,20	0,34	1,07	1,79

Таким образом, выполненные преобразования исходных данных, включающие перераспределение неопределившихся респондентов и увеличение числа точек до необходимого уровня, позволяют перейти к оценке тенденций короткого временного ряда, осуществить теоретическое и экспериментальное исследование возможностей существующих методов прогнозирования.

Теоретический анализ существующих методов прогнозирования на коротких временных рядах

В настоящее время актуальны два подхода для получения прогноза на коротких временных рядах: качественный (экспертные опросы) и количественный [Box et al., 2015]. Так как недостатком качественных методов является слабая надежность [Тамбиева, Попова, Салпагарова, 2015], то с точки зрения обоснованности прогнозов должны применяться либо преимущественно количественные модели, либо комбинация методов.

Прогнозирование на коротких временных рядах принципиально отличается от прогнозирования на временных рядах, имеющих накопленные закономерности развития в прошлом (содержащих более 30 значений), в связи с недостатком априорной информации о поведении ряда. Во-первых, невозможно выявить долгосрочные тенденции развития процесса, так как оценки параметров модели такого ряда ненадежны. Во-вторых, при анализе коротких временных рядов нет возможности использовать сложные модели (например, нейронные сети) для описания трендов, так как их оценивание требует больших непротиворечивых выборок [Домбровский, 2016]. Также в работе [Барбашова, Гайдамакина, Польшакова, 2020] выявлено, что прогнозирование на коротких временных рядах допустимо

DOI: 10.14515/monitoring.2022.2.1921

только на шаг вперед, а применение в прогностических целях полиномов второго и выше порядков недопустимо из-за высокой вероятности кардинального изменения характера динамики исследуемого ряда при экстраполяции полиномиальных моделей за пределами исторических данных. Сравнение существующих количественных методов (моделей) прогнозирования временных рядов проводилось по трем основным показателям (см. табл. 2): требования к исходным данным, срок упреждения, вычислительная сложность [Курилин и др., 2019].

Таблица 2. ***Сравнение методов прогнозирования***

Метод	Показатели			Соответствие
	Требования к исходным данным	Сроки упреждения	Вычислительная сложность	
Экстраполяция тренда	Нет	Краткосрочные	Простой	+
Сглаживание по экспоненте	Нет	Краткосрочные, среднесрочные, долгосрочные	Простой	+
Регрессионные и авторегрессионные модели	Нет	Среднесрочные	Сложный	–
Модель на основе среднего темпа роста	Нет	Краткосрочные	Простой	+
Модель по выборке максимального подобия	Соответствие «подобной» выборке	Среднесрочные	Простой	–

Из таблицы 2 следует, что для прогнозирования на коротких временных рядах будут использоваться следующие количественные модели прогнозирования: модель экспоненциального сглаживания, среднего темпа роста, а также модель экстраполяции аппроксимированной функции временного ряда. Более подробное описание методов представлено ниже.

Суть метода экспоненциального сглаживания заключается в том, что прогноз ожидаемых величин определяется путем взвешенных средних величин текущего периода и сглаженных значений предшествующего рассчитывается по рекуррентной формуле:

$$S_t = \alpha x_t + (1 - \alpha) S_{t-1}, \quad (6)$$

где S_t — значение экспоненциальной средней в момент t; α — параметр сглаживания ($\alpha = \text{const}$; $0 < \alpha < 1$); x_t — фактическое значение исследуемого показателя за период, предшествующий прогнозному. Таким образом, величина S_t является взвешенной суммой всех членов ряда, причем веса падают экспоненциально в зависимости от давности наблюдения. Однако сложности возникают с выбором начального значения S_0 и значением параметра сглаживания.

Метод среднего темпа роста T характеризует отношение двух сравниваемых значений ряда и выражается в процентах:

DOI: 10.14515/monitoring.2022.2.1921

$$T_t = \frac{y_t}{y_{t-1}} \cdot 100\%, \quad (7)$$

где y_t и y_{t-1} — соответствующие два последовательных уровня временного ряда.

Средний темп роста — обобщающая характеристика динамики процесса, отражающая интенсивность изменения уровней ряда. Он показывает, сколько в среднем процентов последующий уровень составляет от предыдущего на всем периоде наблюдений. Этот показатель рассчитывается по формуле средней геометрической из цепных темпов роста:

$$T^{\text{ср}} = \sqrt[n-1]{T_1 T_2 \ldots T_n}, \quad (8)$$

где T_1, T_2, … T_n — цепные темпы роста.

Метод регрессионного анализа основан на аппроксимации временного ряда, то есть замене одних объектов другими, близкими к исходным, но более простыми. Полученная эмпирическая формула обычно справедлива только для узкого интервала измерений. В общем случае регрессия может быть представлена в виде суммы неслучайной и случайной составляющих.

$$y = f(x) + \text{v}, \quad (9)$$

где $f(x)$ — функция регрессионной зависимости, а v — аддитивная случайная величина с нулевым математическим ожиданием.

Адекватность полученного уравнения регрессии определяется по критерию Фишера:

$$F = s^2_{\text{ад}} \Big/ s^2_{\text{восп}}, \quad (10)$$

где $s^2_{\text{ад}}$ — дисперсия адекватности, $s^2_{\text{восп}}$ — дисперсия воспроизводимости [Блохин, 2002]. Уравнение можно считать адекватным эксперименту, если вычисленное значение F окажется меньше для уровня значимости p табличной величины $F_{1-p}(f_{\text{ад}}, f_{\text{восп}})$, где f — соответствующие числа степеней свободы. Повышения адекватности уравнения регрессии эксперименту обычно добиваются увеличением степени аппроксимирующего полинома. Однако при малых объемах выборки увеличение порядка полинома может иногда приводить к росту остаточной дисперсии. Чтобы избежать этого, при решении многих задач производят замену переменных. Например, зависимости типа $\hat{z} = a_0 t^{a_1}$ сводятся к линейным $\hat{y} = b_0 + b_1 x$ следующим образом:

$$\ln \hat{z} = \ln a_0 + a_1 \ln t, \quad (11)$$

где $b_0 = \ln a_0$, $b_1 = a_1$, $x = \ln t$ коэффициенты уравнений находятся методом наименьших квадратов [там же].

Результаты экспериментов по использованию методов прогнозирования на коротких временных рядах

С целью сравнения точности прогнозирования с использованием выбранных в работе методов прогнозирования на коротких временных рядах был осуществлен ряд экспериментов.

DOI: 10.14515/monitoring.2022.2.1921

Первый эксперимент проводился для решения задачи выбора метода экстраполяции аппроксимированной функции временного ряда — вида кривой регрессионного уравнения, форма которой соответствует характеру изменения динамического ряда и выбора значения коэффициента сглаживания в методе экспоненциального сглаживания.

Для выбора вида регрессионной кривой было проведено 4 испытания в соответствии с каждой из рассматриваемых функций (линейная, степенная (квадратичная), показательная, логарифмическая).

Количество временных рядов, достаточное для того, чтобы утверждать, что выборка из генеральной совокупности всех временных рядов (5 основных партий по 85 субъектам) будет репрезентативной (остальные заявленные партии не учитываются, так как процент проголосовавших за них меньше 1), было рассчитано исходя из формулы (12):

$$n = \frac{Z^2 pq \,/\, del^2}{1 + \frac{Z^2 pq \,/\, del^{2-1}}{N}}, \qquad (12)$$

где Z — коэффициент, зависящий от доверительной вероятности (для заданной 95-процентной доверительной вероятности $Z = 1{,}96$), N — объем генеральной совокупности, p ($q = 1 - p$) — доля временных рядов, у которых исследуемый признак присутствует (значения p и q принимаются равными 0,5, поскольку точно неизвестны до проведения исследования, однако при этом значении размер ошибки выборки максимален), del — предельная ошибка выборки (принимается равной 10 %), n — объем выборки.

Таким образом, результаты эксперимента, отражающие точность прогнозного значения, выраженную в частоте встречаемости ошибок при прогнозировании на 75 временных рядах (по 5 партиям в 15 регионах), представлены в таблице 3.

Порог для критерия проверки точности был взят в 5 % (средней точности), для более достоверного прогноза принят порог 3 % (точный) [Баскакова, 2018].

Таблица 3. ***Результаты экспериментов по выбору вида кривой регрессионного уравнения***

Свойство	Критерий	Регрессионная функция, %			
		Линейный	Степенной	Показательный	Логарифмический
Точность	Менее 3 %	24	27	32	40
	3 % — 5 %	39	38	36	32
	Более 5 %	37	35	32	28

Так как прогноз, основанный на логарифмической функции, по сравнению с остальными чаще дает точный результат (72 % временных рядов в совокупности позволили получить результат средней точности), то можно сделать вывод, что

DOI: 10.14515/monitoring.2022.2.1921

аппроксимация логарифмической функцией более предпочтительна для прогнозирования результатов социологических опросов по выборной тематике.

Для выбора коэффициента сглаживания были проведены расчеты с использованием ранее рассмотренных 75 временных рядов. С шагом 0,1 рассчитаны средняя и максимальная ошибки метода (см. табл. 4).

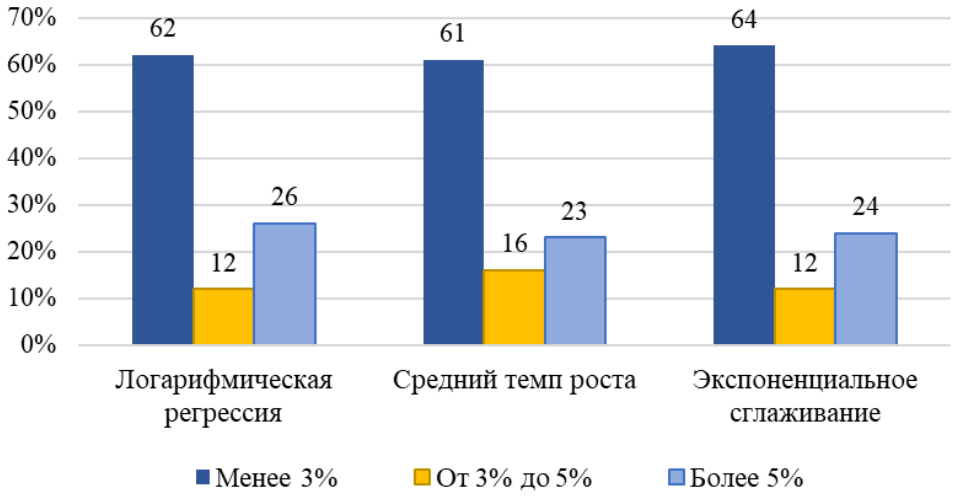
Таблица 4. ***Сравнение точности результатов прогнозов с различной степенью сглаживания***

Коэффициент сглаживания	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1
Средняя ошибка (%)	2,93	2,91	2,86	2,91	2,92	2,95	3	3,2	3,6
Максимальная ошибка (%)	11,72	11,66	11,541	11,55	12,3	13,4	15,1	17,8	22,0

В результате расчетов получено, что средняя и максимальная ошибки минимальны при коэффициенте сглаживания, равном 0,7.

Второй эксперимент заключался в сравнении выбранных ранее методов прогнозирования и выборе лучшего из них по точности прогноза. Критерием точности являлась не превышающая 3% ошибка, вычисленная как разность прогнозного и реального значений. Так, для сформированных исходных данных (см. табл. 1) были получены прогнозные оценки с помощью аппроксимации логарифмической функцией, метода экспоненциального сглаживания и среднего темпа роста, а затем рассчитаны ошибки прогнозирования с использованием фактических данных по результатам выборов, представленные на рисунке 2.

Рис. 2. Частотное распределение ошибок, полученных в результате прогнозирования временных рядов с помощью различных методов



DOI: 10.14515/monitoring.2022.2.1921

Анализ результатов экспериментов по выбору метода прогнозирования показал, что все из представленных методов обеспечивают заданную точность прогнозирования примерно в 60 % случаев. Примерно в 20 % случаев ошибка прогнозирования превышает 5 %, что потребует расширения исследования. Также в результате анализа выбранных в работе методов — экспоненциального сглаживания, среднего темпа роста, экстраполяции аппроксимированной функции временного ряда — выявлено, что при формировании прогнозных показателей участников избирательных кампаний необходимо в дальнейшем учитывать значения суммарного прогноза по каждому участнику, которые не будут сходиться к 100 %, и на заключительном этапе прогнозирования потребуется их корректировка.

В работе была выдвинута гипотеза о том, что повышение точности прогнозирования результатов социологических опросов можно достичь за счет использования аппроксимации логарифмической функцией, а затем корректировки прогнозных значений на основе имеющихся данных, полученных в ходе прошлых избирательных кампаний. Использование результатов выборов прошлых лет возможно из-за выявленного в работе [Чучуева, 2010] свойства подобия двух выборок. Одинаковый фактический результат голосования в большинстве случаев определяет похожесть электоральных ситуаций в регионе. Так как каждое уравнение регрессионной кривой уникально описывает ее расположение относительно системы координат, то необходимо найти такую электоральную ситуацию в прошлом, которая была бы близка к прогнозируемой.

На первом шаге для временного ряда каждой партии строится логарифмический тренд. Но из-за того, что одно и то же уравнение может описывать различное расположение точек относительно начала координат, необходимо выявить, насколько точно логарифмическая функция задает значения, полученные на практике. Оценку соответствия регрессионной модели исходным данным можно произвести на основе ошибки MAPE [Елисеева, 2003]. В качестве критерия близости аппроксимирующей функции к совокупности точек используются модули относительной разности табличных значений y_i и теоретических, рассчитанных по уравнению регрессии $\hat{y}_i$. Следовательно, средняя абсолютная процентная ошибка аппроксимации будет рассчитываться по формуле:

$$MAPE = \frac{1}{n}\sum\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\,\%, \qquad (13)$$

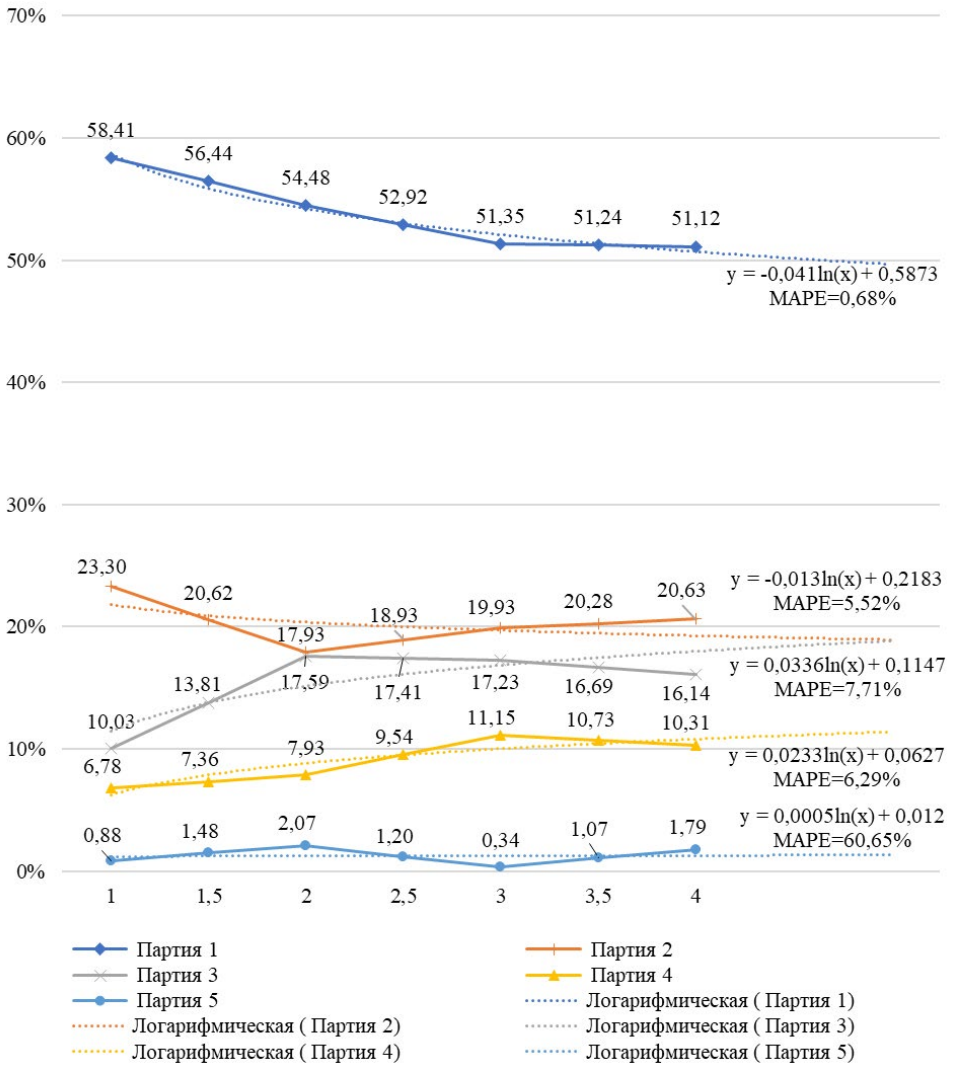
Критерием для проверки MAPE являются следующие значения: до 5 % ошибки — отличный прогноз, от 5 % до 10 % — приемлемый, более 10 % — недопустимый [там же]. Таким образом, для каждого временного ряда получено логарифмическое уравнение и ошибка MAPE (см. рис. 3).

Так как в полученных на рисунке 3 расчетах кроме партии 5 у всех временных рядов MAPE меньше 10 %, то аппроксимированные логарифмической функцией кривые являются приемлемыми. Следовательно, их можно использовать для прогнозирования. Однако для временных рядов, у которых MAPE больше 10 %, следует использовать другой метод прогнозирования. На данных второго эксперимента (см. табл. 5) для логарифмической регрессии была рассчитана ошибка

DOI: 10.14515/monitoring.2022.2.1921

MAPE и выявлено, что в таких случаях наименьшую ошибку прогнозирования позволяет получить метод среднего темпа роста.

Рис. 3. Построение логарифмических уравнений и расчет ошибки MAPE



Второй этап заключается в нахождении похожей электоральной ситуации в прошлом. Такая операция возможна в том случае, когда имеется большой накопленный объем данных, представляющих собой результаты социологических опросов в разных регионах с различными электоральными ситуациями. Предложенная методика предполагает, что электоральная ситуация с близкими к текущим значениями социологических опросов в прошлом существует и коэффициенты, приме-

DOI: 10.14515/monitoring.2022.2.1921

няемые для расчетов прогнозных значений в прошлом так же применимы. Данное допущение ограничивает рассматриваемую методику.

Таблица 5. ***Результаты экспериментов, где MAPE превышает 10 %***

№	MAPE, %	Ошибка прогнозирования, полученная с помощью различных методов, в %		
		Логарифмическая регрессия	Средний темп роста	Экспоненциальное сглаживание
1	60,65	0,09	0,58	0,23
2	32,26	0,83	**0,01**	0,10
3	15,65	3,19	**1,21**	1,54
4	14,47	0,28	**0,22**	0,30
5	14,17	0,36	1,75	1,95
6	14,52	1,62	**0,42**	0,91

Чтобы однозначно описать сложившуюся электоральную ситуацию, необходимо сформировать вектор значений, состоящий из коэффициентов аппроксимированных уравнений для партий по каждому субъекту. Например, в рассмотренном выше примере (см. рис. 3) вектор значений будет выглядеть следующим образом:

(–0,041; 0,5873; –0,013; 0,2183; 0,0336; 0,1147; 0,0233; 0,0627),

где первые два значения вектора соответствуют коэффициенту при логарифме и свободному члену логарифмического уравнения первой партии, вторые два и последующие — к соответствующим партиям в порядке убывания среднего значения по ряду. Если в каких-либо случаях партии меняются местами, то следует учитывать их порядок, меняя пары коэффициентов местами.

Для нахождения похожей ситуации необходимо провести иерархический кластерный анализ [Everitt, 2001], на первом этапе которого будет найдено самое ближайшее «расстояние» до сформированного n-мерного вектора (предварительно по данным социологических опросов за предыдущие года также были сформированы вектора) с помощью метрики «Евклидово расстояние».

Однако из-за того, что регрессионное уравнение не всегда точно описывает расположение точек относительно начала координат, необходимо сравнить MAPE каждой из выбранных кривых. Если ошибка MAPE больше 10 %, то для прогнозирования нельзя использовать эту электоральную ситуацию, и необходимо рассмотреть второй «по близости» вектор. Итерации повторяются до тех пор, пока не будет найден удовлетворяющий всем условиям вектор.

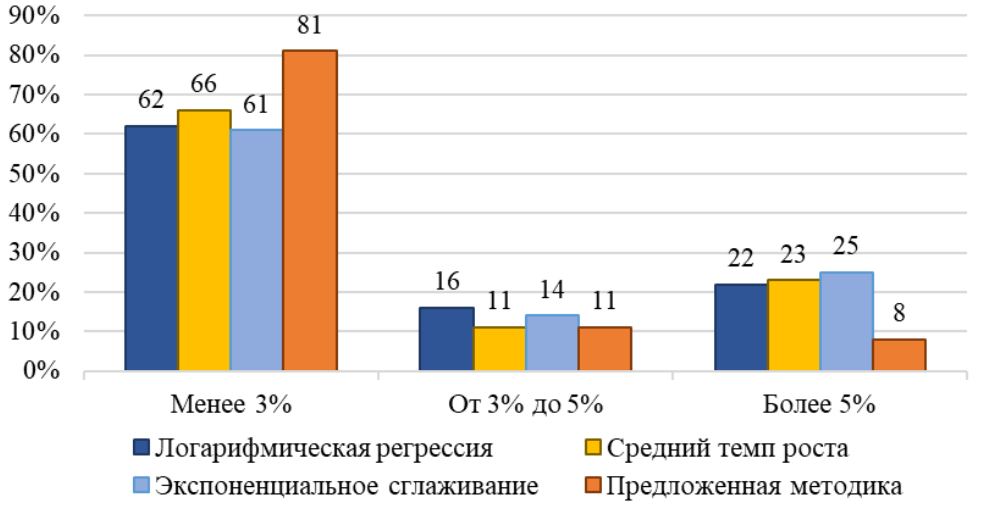
Для корректировки полученных прогнозных значений необходимо учитывать изменение фактических значений в «исторически сложившейся» электоральной ситуации по сравнению с результатами последнего социологического опроса. Под «исторически сложившейся» электоральной ситуацией следует понимать совокупность логарифмических кривых и точек, описывающих распределение частот,

DOI: 10.14515/monitoring.2022.2.1921

проголосовавших за каждого кандидата в течение проведенных социологических опросов, которая найдена кластерным анализом по ближайшему вектору. Если наблюдается увеличение числа проголосовавших за первого кандидата, то следует сделать вывод, что, скорее всего, и в текущей электоральной ситуации процент проголосовавших увеличится. Следовательно, необходимо скорректировать прогноз по всем кандидатам, изменив значение с учетом поправочной разницы. Для приведения общей суммы к 100 % необходимо пропорционально увеличить (уменьшить) «остаток», являющийся разницей суммы полученных с помощью различных методов прогнозных оценок и 100 %. Таким образом, по каждой партии были получены прогнозные оценки, значительно не отклоняющиеся от фактических значений и соответствующие динамике изменения мнения общества под влиянием различных общественно-политических и социально-экономических процессов.

Третий эксперимент проводился с целью проверки гипотезы, связанной с повышением точности прогнозирования за счет использования данных, полученных в ходе прошлых избирательных кампаний. Эксперимент заключался в формировании прогнозов результатов выборов с помощью предложенной и существующих методик на десяти случайно выбранных контрольных выборках в субъектах РФ. В качестве показателя точности использовалось отклонение полученного прогноза от фактического результата (если разница не превышала 3 %, то прогноз считается высокой точности, если от 3 % до 5 %, то средней точности, если более 5 %, то низкой точности). Точность логарифмического приближения также предлагалось оценивать с использованием выражения (13). Сравнение результатов, полученных с помощью предложенной методики и существующего метода прогнозирования, проводилось по критерию превосходства. Частота получения ошибки в прогнозах с помощью различных методов по партиям представлена на рисунке 4.

Рис. 4. Частота встречаемости различных ошибок прогнозирования в зависимости от используемого метода



DOI: 10.14515/monitoring.2022.2.1921

В результате эксперимента была подтверждена гипотеза о возможности повышения точности прогнозирования с использованием предложенных процедур, которые были оформлены в виде методики прогнозирования на коротких временных рядах и графическое представление которой представлено на рисунке 5.

Рис. 5. Методика прогнозирования на коротких временных рядах

Для возможности автоматизации предложенных процедур в методике была разработана программа для ЭВМ [2].

Заключение

Прогнозирование результатов выборов на основе социологических опросов является актуальной задачей и может решаться различными методами. Однако существующие методы не позволяют получить высокую точность прогнозов на основе коротких временных рядов. В настоящей работе были использованы метод экспоненциального сглаживания, среднего темпа роста и экстраполяции временного ряда. Для метода экстраполяции в качестве аппроксимирующей функции была взята логарифмическая — в связи с тем, что она дает наиболее точный прогноз. Для метода экспоненциального сглаживания коэффициент сглаживания был выбран равный 0,7, так как в результате эксперимента при этом значении коэффициента получается наименьшая средняя и максимальная ошибка.

[2] Воробьев А. А. Программный модуль прогнозирования результатов избирательных кампаний на коротких временных рядах. Свидетельство о регистрации программы для ЭВМ RU2020666538, 11.12.2021. Заявка № 2020665715 от 03.12.2020.

DOI: 10.14515/monitoring.2022.2.1921

В результате анализа проведенных в работе экспериментов выяснилось, что все рассматриваемые методы с одинаковой точностью дают прогноз, однако примерно 20 % всех прогнозов являются «неточными» (ошибка превышает 5 %).

В связи с этим была разработана и апробирована методика, заключающаяся в аппроксимации временного ряда логарифмическим уравнением, построением вектора, описывающего электоральную ситуацию, нахождением методом иерархической кластеризации наиболее близких социологических данных прошлых лет, а затем пропорционального перераспределения «остатка», являющегося разницей суммы полученных с помощью метода экстраполяции прогнозных оценок и 100 %.

Применение предложенной методики прогнозирования обеспечивает повышение точности по сравнению с существующими методами на рассмотренных исходных данных. За счет сравнения текущей и «исторически сложившейся» электоральной ситуации найдены значения, позволяющие скорректировать и перераспределить «остаток», уменьшив таким образом количество партий, на основе которых был получен неточный прогноз в исследуемых регионах.

Список литературы (References)

Барбашова Е. В., Гайдамакина И. В., Польшакова Н. В. Прогнозирование в коротких временных рядах: методологические и методические аспекты // Вестник аграрной науки. 2020. № 2. С. 84—98. https://doi.org/10.17238/issn2587-666X.2020.2.84.
Barbashova E. V., Gaydamakina I. V., Polshakova N. V. (2020) Forecasting in Short Time Series: Methodological and Methodical Aspects. *Bulletin of Agrarian Science*. No. 2. P. 84—98. https://doi.org/10.17238/issn2587-666X.2020.2.84. (In Russ.)

Баскакова Ю. М. Сравнительная оценка точности разных методов прогнозирования результатов выборов // Материалы VIII Международной социологической Грушинской конференции «Социолог 2.0: трансформация профессии», 18—19 апреля 2018 г. URL: https://profi.wciom.ru/index.php?id=1771 (дата обращения: 13.04.2022).
Baskakova Yu.M. (2018) Comparative Assessment of the Accuracy of Different Methods for Predicting Election Results. In: *Proceedings of the VIII International Sociological Grushin Conference "Sociologist 2.0: Transformation of the Profession"*, April 18—19, 2018. URL: https://profi.wciom.ru/index.php?id=1771 (accessed: 13.04.2022). (In Russ.)

Блохин А. В. Теория эксперимента: курс лекций в двух частях. Минск : Научно-методический центр «Электронная книга БГУ», 2002.
Blokhin A.V. (2002) Theory of Experiment: a Course of Lectures in Two Parts. Minsk: Scientific and Methodological Center "Electronic Book of Belarusian State University". (In Russ.)

Воробьев А. А., Воронецкий А. А. Исследование возможностей математических методов по перераспределению неопределившихся респондентов // Мягкие измерения и вычисления. 2019. № 9. С. 35—39.
Vorobiev A. A., Voroneckiy A. A. (2019) Research of Possibilities of Mathematical Methods for Redistributing Uncertain Respondents. *Soft Measurements and Calculations*. No. 9. P. 35—39. (In Russ.)

DOI: 10.14515/monitoring.2022.2.1921

Воробьев А. А., Воронецкий А. А., Азрапкин А. И., Белоножко Е. Д. Исследование возможностей математических методов по восстановлению пропусков в номинативных социологических данных // Системы управления и информационные технологии. 2020. № 2. С. 93—97.
Vorobyov A. A., Voronetsky A. A., Azrapkin A. I., Belonozhko E. D. (2020) Research of Possibilities of the Mathematical Methods for Restoring Omissions in Nominative Sociological Data. *Journal of Control Systems and Information Technology*. No. 2. P. 93—97. (In Russ.)

Горшков М. К., Шереги Ф. Э. Прикладная социология: методология и методы. М.: Институт социологии РАН. 2011. С. 270—275.
Gorshkov M. K., Sheregi F. E. (2011) Applied Sociology: Methodology and Methods. Moscow: Institute of Sociology RAS. P. 270—275. (In Russ.)

Домбровский В. В. Анализ и прогнозирование коротких временных рядов с привлечением экспертной информации // Эконометрика. Томск: Томский государственный университет, 2016. URL: https://lib.tsu.ru/mminfo/2016/Dombrovski/book/chapter-8/chapter-8.htm (дата обращения: 28.03.2021).
Dombrowski V. V. (2016) Analysis and Forecasting of Short Time Series with the Involvement of Expert Information. In: *Econometrics*. Tomsk: Tomsk State University. URL: https://lib.tsu.ru/mminfo/2016/Dombrovski/book/chapter-8/chapter-8.htm (accessed: 28.03.2021) (In Russ.)

Елисеева И. И. Эконометрика. М.: Финансы и статистика. 2003. С. 225—262.
Eliseeva I. I. (2003) Econometrics. Moscow: Finance and Statistics. P. 225—262. (In Russ.)

Жучкова С. В., Ротмистров А. Н. Возможность работы с пропущенными данными при использовании CHAID: результаты статистического эксперимента // Социология: 4М. 2018. № 46. С. 85—122.
Zhuchkova S. V., Rotmistrov A. N. (2018) Handling Missing Data with CHAID: Results of a Statistical Experiment. *Sociology: 4M*. No. 46. P. 85—122. (In Russ.)

Зангиева И. К. Проблема пропусков в социологических данных: смысл и подходы к решению // Социология: 4М. 2011. № 33. С. 28—56.
Zangieva I. K. (2011) The Problem of Missing Values in Sociological Data: Essence and Solution Methods. *Sociology: 4M.* No. 33. P. 28—56. (In Russ.)

Знаменский С. В. Численная оценка точности интерполяции несложных элементарных функций // Программные системы: теория и приложения. 2018. № 4. С. 69—92. https://doi.org/10.25209/2079-3316-2018-9-4-69-92.
Znamenskij S. V. (2018) Numerical Evaluation of the Interpolation Accuracy of Simple Elementary Functions. *Program Systems: Theory and Applications*. No. 4. P. 69—92. https://doi.org/10.25209/2079-3316-2018-9-4-69-92. (In Russ.)

Капитанова О. В. Прогнозирование социально-экономических процессов. Нижний Новгород: Нижегородский госуниверситет, 2016.
Kapitanova O. V. (2016) Forecasting Socio-economic Processes: Study Guide. Nizhny Novgorod: Nizhny Novgorod State University. (In Russ.)

DOI: 10.14515/monitoring.2022.2.1921

Клисторин В. И. О точности и надежности прогнозов // ЭКО. 2011. № 12. С. 40—47.
Klistorin V. I. (2011) About the Accuracy and Reliability of Forecasts. *ECO Journal*. No. 12. P. 40—47. (In Russ.)

Курилин Б. Л., Киселевская-Бабинина В. Я., Карасёв Н. А., Киселевская-Бабинина И. В., Кислухина Е. В., Васильев В. А. Выбор метода прогнозирования основных статистических показателей работы ГБУЗ «НИИ СП им. Н. В. Склифосовского Департамента здравоохранения города Москвы» // Журнал им. Н. В. Склифосовского «Неотложная медицинская помощь». 2019. Т. 8. № 3. С. 246—256. https://doi.org/10.23934/2223-9022-2019-8-3-246-256.
Kurilin B. L., Kiselevska-Babinina V.G., Karasv N. A., Kiselevska-Babinina I.V., Kislukhina E. V., Vasilyev V. A. (2019) Selection of Prediction Method of Basic Statistical Work Parameters of N. V. Sklifosovsky Research Institute for Emergency Medicine of the Moscow Healthcare Department. *Russian Sklifosovsky Journal "Emergency Medical Care"*. Vol. 8. No. 3. P. 246—256. https://doi.org/10.23934/2223-9022-2019-8-3-246-256. (In Russ.)

Тамбиева Д. А., Попова Е. В., Салпагарова Ш. Х. К проблеме недостаточности информации. Малые выборки или «очень короткие» временные ряды // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2015. № 107. С. 126—141.
Tambieva D. A., Popova E. V., Salpagarova Sh.H. (2015) To the Problem of Lack of Information. Small Samples or "Very Short" Time Series. *Polythematic Online Scientific Journal of Kuban State Agrarian University*. No. 107. P. 126—141. (In Russ.)

Фабрикант М. С. Модель-ориентированный подход к отсутствующим значениям: множественная импутация в многоуровневой регрессии посредством R (на примере анализа опросных данных) // Социология: 4М. 2015. № 41. С. 7—29.
Fabrykant M. S. (2015) Model-Oriented Approach to Missing Values: Multiple Imputation in Multilevel Regression Using R (On the Example of Analyzing Survey Data). *Sociology: 4M*. No. 41. P. 7—29. (In Russ.)

Чучуева И. А. Модель прогнозирования временных рядов по выборке максимального подобия // Информационные технологии. 2010. № 12. С. 43—47.
Chuchueva I. A. (2010) Model for Forecasting Time Series on a Sample of Maximum Similarity. *Information Technology*. No. 12. P. 43—47. (In Russ.)

Box G. E.P., Jenkins G. M., Reinsel G. C., Ljung G. M. (2015) Time Series Analysis: Forecasting and Control. 5th Edition Hoboken, NJ: John Wiley & Sons.

Enders C. (2010) Applied Missing Data Analysis. New York, NY: The Guilford Press.

Everitt B. S., Landau S., Leese M. (2001) Miscellaneous Clustering Methods. In: *Cluster Analysis*. New York, NY: Taylor & Francis. P. 141—176.

Zhuchkova S., Rotmistrov A. (2022) How to Choose an Approach to Handling Missing Categorical Data: (Un)expected Findings from a Simulated Statistical Experiment. *Quality and Quantity*. Vol. 56. P. 1—22. https://doi.org/10.1007/s11135-021-01114-w.