Метод «случайный лес» в исследовании влияния макроэкономических показателей регионального развития на уровень неформальной занятости

Елена Викторовна Зарова,
Эльвира Ивановна Дубравская
ГБУ «Аналитический центр» Правительства Москвы, г. Москва, Россия

Тематика количественных исследований неформальной занятости как в Российской Федерации, так и в других странах имеет стабильно высокую и при этом периодически резко возрастающую актуальность, что обусловлено характерностью этого явления для стран с любым уровнем экономического развития и его высокой зависимостью от цикличности и кризисных этапов в их экономической динамике.

В теоретических и прикладных исследованиях особое внимание к оценке факторов и условий неформальной занятости в Российской Федерации связано с необходимостью выработки эффективных мер государственной политики по преодолению негативного влияния неформальной занятости, в том числе на региональном уровне. Среди отрицательных эффектов неформальной занятости, вызывающих обеспокоенность федеральных и региональных органов власти, можно выделить недополучение налогов, потенциальные потери, обусловленные снижением эффективности производства, негативные социальные последствия. Для их преодоления необходима разработка количественных индикаторов, определяющих уровень неформальной занятости в регионах с учетом их специфики в общей пространственно-экономической системе России.

В статье предложены и апробированы методы решения задачи выявления и оценки влияния иерархической взаимосвязи макроэкономических показателей регионального развития на уровень неформальной занятости в субъектах Российской Федерации. Большинство работ, посвященных исследованию неформальной занятости, основано на базовых статистических методах пространственно-динамического анализа, а также на ставших «традиционными» методах кластерного и корреляционно-регрессионного анализа. Не умаляя достоинств этих методов, необходимо отметить их определенную ограниченность в выявлении скрытых структурных связей и взаимозависимостей в таком сложном многомерном явлении, как неформальная занятость.

С целью обоснования возможности преодоления этих ограничений в статье предложены показатели региональной статистики, прямо или косвенно характеризующие неформальную занятость, а также представлены результаты применения метода «случайный лес» для выделения групп субъектов Российской Федерации на основе сходных макроэкономических параметров, определяющих неформальную занятость. Новизна данного метода с точки зрения целей исследования состоит в том, что он позволяет оценить влияние макроэкономических показателей регионального развития на уровень неформальной занятости с учетом неявных, не предопределенных исходными гипотезами иерархических взаимосвязей факторных показателей.

На основе обобщения исследований, представленных в литературных источниках, а также выполнения авторами статистических расчетов с использованием данных Росстата сделаны выводы о высокой значимости макроэкономических параметров регионального развития и системных связей макроэкономических показателей в обосновании дифференциации уровня неформальной занятости по субъектам Российской Федерации.

Ключевые слова: метод «случайный лес», интеллектуальный анализ данных, дерево решений, классификация, регрессия, неформальная занятость, регионализация.

JEL: C38, J21, J46, O18, R11.

doi: https://doi.org/10.34023/2313-6383-2020-27-6-37-55.

Для цитирования: Зарова Е.В., Дубравская Э.И. Метод «случайный лес» в исследовании влияния макроэкономических показателей регионального развития на уровень неформальной занятости. Вопросы статистики. 2020;27(6):37-55.

The Random Forest Method in Research of Impact of Macroeconomic Indicators of Regional Development on Informal Employment Rate

Elena V. Zarova,
Elvira I. Dubravskaya
Analytical Center by Moscow City Government, Moscow, Russia

The topic of quantitative research on informal employment has a consistently high relevance both in the Russian Federation and in other countries due to its high dependence on cyclicality and crisis stages in economic dynamics of countries with any level of economic development.

Developing effective government policy measures to overcome the negative impact of informal employment requires special attention in theoretical and applied research to assessing the factors and conditions of informal employment in the Russian Federation including at the regional level. Such effects of informal employment as a shortfall in taxes, potential losses in production efficiency, and negative social consequences are a concern for the authorities of the federal and regional levels. Development of quantitative indicators to determine the level of informal employment in the regions, taking into account their specifics in the general spatial and economic system of Russia are necessary to overcome these negative effects.

The article proposes and tests methods for solving the problem of assessing the impact of hierarchical relationships on macroeconomic factors at the regional level of informal employment in constituent entities of the Russian Federation. Majority of the works on the study of informal employment are based on basic statistical methods of spatial-dynamic analysis, as well as on the now «traditional» methods of cluster and correlation-regression analysis. Without diminishing the merits of these methods, it should be noted that they are somewhat limited in identifying hidden structural connections and interdependencies in such a complex multidimensional phenomenon as informal employment.

In order to substantiate the possibility of overcoming these limitations, the article proposes indicators of regional statistics that directly and indirectly characterize informal employment and also presents the possibilities of using the «random forest» method to identify groups of constituent entities of the Russian Federation that have similar macroeconomic factors of informal employment. The novelty of this method in terms of research objectives is that it allows one to assess the impact of macroeconomic indicators of regional development on the level of informal employment, taking into account the implicit, not predetermined by the initial hypotheses, hierarchical relationships of factor indicators.

Based on the generalization of the studies presented in the literature, as well as the authors' statistical calculations using Rosstat data, the authors came to the conclusion about the high importance of macroeconomic parameters of regional development and systemic relationships of macroeconomic indicators in substantiating the differentiation of the informal level across the constituent entities of the Russian Federation.

Keywords: random forest method, data mining, decision tree, classification, regression, informal employment, regionalization.
JEL: C38, J21, J46, O18, R11.
doi: https://doi.org/10.34023/2313-6383-2020-27-6-37-55.

For citation: Zarova E.V., Dubravskaya E.I. The Random Forest Method in Research of Impact of Macroeconomic Indicators of Regional Development on Informal Employment Rate. *Voprosy Statistiki*. 2020;27(6):37-55. (In Russ.)

Введение

В исследовании Группы Всемирного банка «Проблема неформальной занятости в России. Причины и варианты решения» [1] отмечено, что в России доля неформально занятых в многолетнем периоде не превышает аналогичный показатель стран, сопоставимых по уровню валового регионального продукта на душу населения, и при этом «...в России неформальная занятость имеет природу, которая отличается от той, что наблюдается в большинстве других стран. Это объясняется тем, что для России характерен высокий уровень образования, экономика не является аграрной, а заняты в ней главным образом наемные работники (а не самозанятые)» [1, с. 11]. Сравнительный анализ общих и специфических факторов неформальной занятости в России на межстрановом уровне является важным направлением статистических исследований. Вместе с тем для выработки эффективной политики легализации трудовых отношений необходима разработка инструментария выявления и оценки значимых региональных факторов, определяющих территориальную дифференциацию уровня неформальной занятости в Российской Федерации. Такой инструментарий должен включать как традиционные методы статистического анализа, так и методы интеллектуального анализа данных (data mining). Эти стремительно развивающиеся

методы позволяют исследовать не только явные, но и скрытые, заранее не предусмотренные в исходных гипотезах связи в изучаемом признаковом пространстве, что наиболее важно для явлений с высокой латентностью статистического наблюдения, к которым относится неформальная занятость.

Актуальность исследования и обзор публикаций

Согласно информации, опубликованной на сайте Международной организации труда (МОТ), более шести работников из десяти и четырех предприятий из пяти в мире заняты в неформальной экономике. Вопреки имеющимся прогнозам, неформальность со временем не уменьшается, а во многих странах даже увеличивается[1]. В российской экономике тенденция роста доли занятых в неформальном секторе имела место с начала 2000-х годов до 2016 г. Как отмечено в упомянутом выше исследовании Группы Всемирного банка, «...этот показатель вырос в первой половине 2000-х, затем на непродолжительное время он стабилизовался и даже снизился, однако впоследствии продолжил рост. Так, по оценкам Росстата, доля занятых в неформальном секторе выросла с 12,5% в 2001 г. до 17,6% к 2005 г. (с небольшим снижением до 16,4% в 2010 г.), а затем последовало дальнейшее существенное ее увеличение, - до 21,1% к 2016 году» [1, с. 11 и 12]. Вместе с тем в последние годы в России наметилась противоположная тенденция: по данным федерального выборочного обследования рабочей силы, с 2016 по 2019 г. доля неформально занятых лиц в общей численности занятых снизилась с 21,2% до 20,6 и ко второму кварталу 2020 г. составила 19,1%. Наблюдаемая в последние годы смена тенденции в динамике неформальной занятости является следствием проводимой государством политики легализации труда самозанятых, развития мер формализации трудовых отношений, повышения уровня экономического развития.

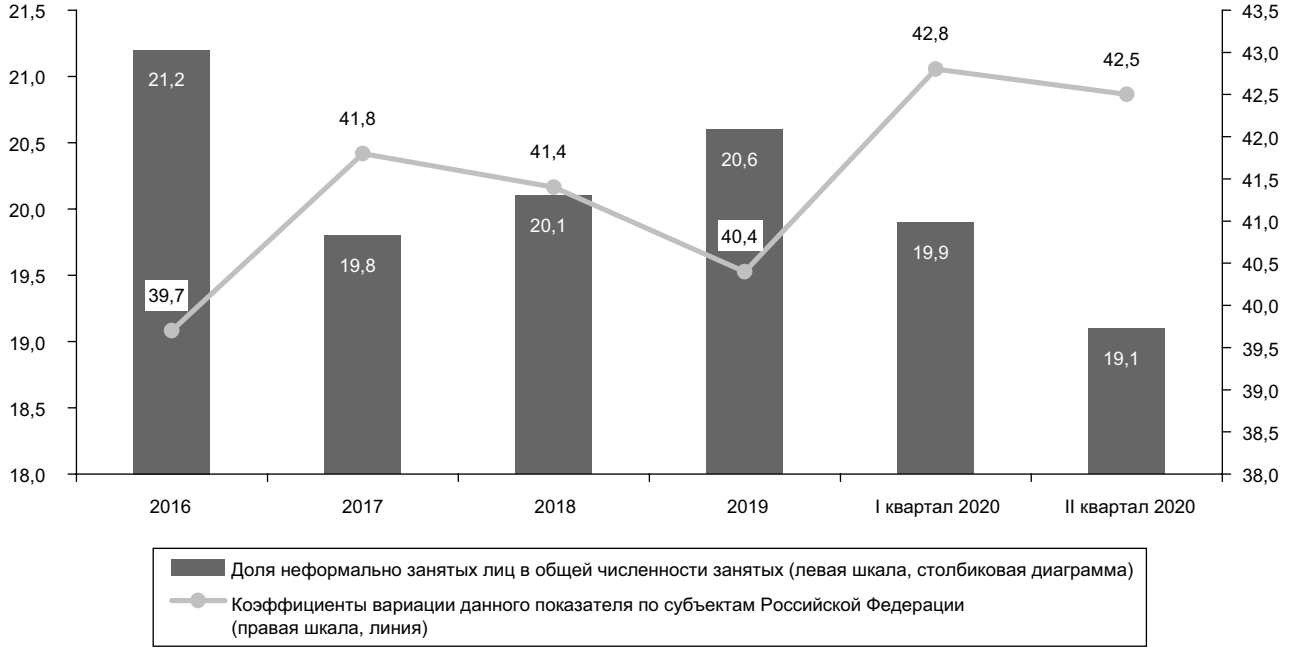


Рис. 1. **Доля неформально занятых лиц в общей численности занятых и коэффициенты вариации данного показателя по субъектам Российской Федерации, 2016 г. - II квартал 2020 г.** (в процентах)

Источник: расчеты авторов по данным Росстата (Итоги выборочного обследования рабочей силы. URL: https://rosstat.gov.ru/compendium/document/13265).

[1] International Labour Organization. Informal economy. URL: http://www.ilo.ch/global/topics/employment-promotion/informal-economy/lang--en/index.htm.

Вместе с тем из данных, представленных на рис. 1, следует, что за рассматриваемый период неоднородность субъектов Российской Федерации по уровню неформальной занятости не снижается: в 2016 г. коэффициент вариации составил 39,7%, в 2019 г. - 40,4, а в 2017 и 2018 гг. его значения превышали 41%.

Статистическое исследование факторов региональной дифференциации уровня неформальной занятости необходимо для выработки информационного обеспечения политики легализации труда, эффективность которой в значительной степени определяется учетом региональной специфики закономерностей неформальной занятости. При этом, как справедливо замечает Дженнифер Э. Коэн, «...формальные и неформальные условия занятости, не только структурно разобщены, но и определяются макроэкономической средой» (перевод наш. - *Е. З.* и *Э. Д.*) [2, p. 14]. Из этого утверждения следует, что необходимым этапом выявления и оценки региональных факторов неформальной занятости должно быть исследование количественных закономерностей зависимости неформальной занятости от макроэкономических показателей регионального развития.

Определение неформальной занятости

Согласно рекомендациям МОТ, лица имеют неформальную работу, если их трудовые отношения по закону или на практике не подпадают под действие национального трудового законодательства, налогообложения доходов, социальной защиты или права на определенные пособия по найму (например, предварительное уведомление об увольнении, выходное пособие, оплачиваемый ежегодный отпуск или отпуск по болезни и т. д.). Основные причины неформальной занятости объединены в следующие группы: недекларирование места работы, случайная работа или работа на ограниченный короткий срок; рабочие места с часами работы или заработной платой ниже установленного порога; занятость лиц на некорпорированных предприятиях или в домашних хозяйствах.

В Резолюции о статистике занятости в неформальном секторе, принятой 15-й Международной конференцией статистиков труда (МКСТ, январь 1993 г.)[2], а также в Руководящих принципах, касающихся статистического определения неформальной занятости, утвержденных на 17-й МКСТ (ноябрь 2003 г.)[3], установлено, что к работникам в составе неформальной занятости относятся:

1. Работники неформального сектора (за исключением работников, с которыми заключен контракт на предприятиях неформального сектора).

2. Работники за пределами неформального сектора, в том числе:
- работники, выполняющие неформальную работу на предприятиях формального сектора (НФС);
- семейные работники, помогающие работникам НФС;
- оплачиваемые домашние работники, нанятые домашними хозяйствами на неформальную работу;
- работники, работающие на индивидуальной основе, занятые производством товаров исключительно для собственного конечного использования в своем домашнем хозяйстве.

Таблица 1

Состав неформальной занятости

Производственные единицы	Неформальные трудовые отношения	Формальные трудовые отношения
Единицы неформального сектора (индивидуальные предприниматели, зарегистрированные самозанятые)	A	B
Другие производственные единицы (в том числе организации)	C	D

Источник: [3].

Согласно схеме, представленной в таблице 1, категории работников по типу неформальной занятости, содержащиеся в публикациях Росстата, систематизированы следующим образом[4]:

A + C = лица, имеющие неформальную занятость («неформальная занятость»);

A + B = лица, занятые в неформальном секторе;

C = неформальная занятость за пределами неформального сектора;

B = формальная занятость в неформальном секторе.

[2] URL: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/normativeinstrument/wcms_234473.pdf.
[3] URL: https://www.ilo.org/public/english/bureau/stat/download/guidelines/russian/defempl.pdf.
[4] Теоретическое обоснование представленной систематизации приведено в источнике [3].

Для целей количественной оценки неформальной занятости, измеряемой в соответствии с системой категорий, представленных в таблице 1, использованы годовые данные федерального выборочного обследования рабочей силы в разрезе субъектов Российской Федерации, опубликованные Росстатом[5].

Обзор публикаций по теме исследования

Количественные исследования макроэкономических факторов неформальной занятости описаны в ряде публикаций. Примеры этих работ приведены ниже, что дает представление об аналитических задачах, решаемых авторами, и применяемых ими статистических методах.

В работе [3] с использованием методов построения и анализа панельной регрессии исследуется обусловленность уровня неформальной занятости в городах Китая показателями, характеризующими темпы и этапы их экономического развития, выпуск в отраслях, производящих продукты и услуги, безработицу, миграцию из сельских районов в города и глобализацию городской экономики. Моделирование на панельных данных применяется и в работе [4] для оценки влияния макроэкономических переменных на объем неформальной экономики в 98 развивающихся странах. При этом в состав факторных переменных автором включены индексы инфляции и безработицы, показатели, характеризующие открытость экономики и уровень коррупции, долю нефтедобывающих отраслей в ВВП.

Также на основе применения методов панельной регрессии в работе [5] исследовано влияние затрат на труд на уровень неформальной занятости в странах ОЭСР и Латинской Америки. При этом выявлена специфика закономерностей этого влияния для двух групп стран, отличающихся типом экономического роста:

1) обусловленного фактором «вознаграждение за труд» (*wage-led growth*) - для него характерна акселерационная роль повышения доли оплаты труда в совокупном доходе;

2) обусловленного фактором «прибыль» (*profit-led growth*) - для него основным стимулом служит увеличение доли прибыли[6].

Наряду с теоретическими исследованиями, в литературе представлены прикладные решения международных организаций по измерению макроэкономических факторов неформальной занятости на страновом и региональном уровнях. В качестве примера можно привести анализ группировки 36 стран по характеристикам зависимости уровня неформальной занятости от показателей экономического роста, безработицы, уровня бедности, опубликованный по результатам исследования, проведенного специалистами Департамента статистики МОТ [7]. По итогам этого анализа сделан следующий вывод: «Неформальная занятость отрицательно коррелирует с доходом на душу населения и положительно коррелирует с бедностью в разных странах. Это говорит о том, что по мере роста ВВП и (или) сокращения бедности в разных странах работники с большей вероятностью будут знать о своих правах на определенные юридические и социальные гарантии и льготы и успешно получат такую защиту и льготы» (перевод наш. - *Е. З.* и *Э. Д.*) [7, p. 2].

Представление о значимых макроэкономических показателях развития, влияющих на уровень неформальной занятости, дают также опубликованные результаты проведенного Центром развития ОЭСР исследования «Устранение уязвимости в неформальной экономике», в котором отмечается: «Анализ макроданных проливает свет на связь между неформальностью и развитием. С учетом ряда факторов распространенность неформальной экономики в разных странах коррелирует с ключевыми результатами развития. Эконометрический анализ обнаруживает отрицательную связь между уровнем неформальной занятости и валовым внутренним продуктом (ВВП), индексами человеческого развития и производительности труда, а также положительную корреляцию с уровнем бедности» (перевод наш. - *Е. З.* и *Э. Д.*) [8, p. 17].

Обобщение представленных выше публикаций подтверждает актуальность и теоретическую значимость статистического исследования влияния макроэкономических показателей регионального развития на уровень неформальной занятости, является основой для выработки новых алгоритмов решения этой задачи с применением методологических подходов интеллектуального анализа данных (data mining).

[5] Росстат. Итоги выборочного обследования рабочей силы. URL: https://rosstat.gov.ru/compendium.

[6] Подробное изложение теоретических положений двух экономических типов - «wage-led-growth» и «profit-wage-growth» - представлено в книге [6].

Постановка задачи и исходная информационная база

Для решения задачи выявления статистически однородных групп субъектов Российской Федерации на основе количественных характеристик влияния макроэкономических показателей социально-экономического развития на уровень неформальной занятости сформирована система статистических индикаторов, включающая результативный показатель - численность неформально занятых лиц в процентах к общей численности занятого населения - и факторные показатели. Последние характеризуют социально-экономическое развитие субъектов Российской Федерации (см. таблицу 2). Факторные показатели содержат макроэкономические оценки уровня и темпов экономического развития субъектов Российской Федерации, основанные на данных системы национальных счетов регионального уровня, показатели инфляции, безработицы, реальных доходов населения, оплаты труда в регионах. В указанной таблице представлены характеристики лагов запаздывающего влияния факторных показателей на результативный показатель. Периоды лагов запаздывающего влияния факторных показателей определяются максимальным значением коэффициентов парной корреляции с зависимой величиной при соответствующих сдвигах во времени связных временных рядов. В ряде случаев включенные в модель лаговые эффекты определяются отсутствием значений факторных показателей на последнюю точку анализируемых временных рядов (2020 г.), что характерно для показателей, основанных на данных о валовом региональном продукте.

Таблица 2

Показатели, используемые для выявления и оценки количественных закономерностей влияния макроэкономических показателей социально-экономического развития субъектов Российской Федерации на уровень неформальной занятости

Показатели	Условное обозначение	Лаг запаздывающего влияния, лет
Результативный показатель - доля неформальной занятости (удельный вес лиц, имеющих неформальную занятость в неформальном и формальном секторах экономики в общей численности занятых), в процентах	K6	_
Факторные показатели, в том числе:		
Индекс физического объема валового регионального продукта (ВРП) на душу населения, в процентах	A_18	1
Валовой региональный продукт на душу населения, рублей	B_18	1
Доля валового регионального продукта субъекта Российской Федерации в валовом региональном продукте Российской Федерации, в процентах	C_18	1
Отношение объема инвестиций в основной капитал к ВРП, в процентах	D_18	1
Индекс производительности труда (по методологии СНС), в процентах	E_18	1
Доля продукции высокотехнологичных и наукоемких отраслей в ВРП, в процентах	F_18	1
Доля работников микропредприятий в общей среднесписочной численности, в процентах	G_18	1
Доля работников малых предприятий в общей среднесписочной численности, в процентах	H_18	1
Фактическое конечное потребление домашних хозяйств в расчете на 1 человека на территории субъектов Российской Федерации (в текущих рыночных ценах), рублей	I_17	2
Среднемесячная заработная плата работника полного круга организаций, рублей	ZP_19	0
Среднемесячный трудовой доход (среднемесячная номинальная начисленная заработная плата наемных работников организаций, у индивидульных предпринимателей и физических лиц), рублей	ZNR_19	0
Уровень бедности, в процентах	UB_19	0
Индекс потребительских цен, в процентах	IPC_18	1
Реальные денежные дохода населения, в процентах	RDdox_18	1
Отраслевая структура ВРП (доля вида деятельности), в процентах	SA...ST_18	1

Примечания: а) лаги запаздывающего влияния (лет) указаны по отношению к значениям результативного показателя за 2019 г.; б) в качестве исходных данных по приведенным показателям в разрезе субъектов Российской Федерации использовались данные, опубликованные на официальном сайте Росстата (https://rosstat.gov.ru), и данные ЕМИСС (https://www.fedstat.ru).

Источник: расчеты авторов по данным Росстата (Регионы России. Социально-экономические показатели. URL: https://rosstat.gov.ru/folder/210/document/13204).

Как было представлено выше в обзоре литературы, во многих работах для решения аналогичных задач применяется построение регрессионных моделей на панельных данных, спецификация которых основана на линейных взаимосвязях результативного и факторных показателей.

Линейные регрессионные модели для решения обратной задачи - анализа зависимости производительности труда от преобладающего типа занятости (формальной/неформальной) - представлены в работе «Неформальный сектор во франкоязычных странах Африки. Размер фирм, производительность и институты» [9], подготовленной специалистами Всемирного банка. При этом авторы, оценивая полученные результаты, делают верные замечания о некоторых условиях, которые могли повлиять на качество регрессионных моделей, в том числе следующие:

1. Большинство переменных не имеют нормального распределения, и многие из них имеют существенную асимметрию распределения.

2. Нелинейная спецификация может дать лучшие результаты.

3. Отрицательная корреляция между уровнями неформальной занятости и производительности не обязательно указывает направление причинной связи; это может быть результатом двунаправленной причинно-следственной зависимости.

Отмечая справедливость приведенных замечаний, необходимо обратить внимание на то, что методы многофакторного регрессионного моделирования основаны на заранее сформированных гипотезах о наборе факторных переменных и характере их статистической связи с зависимой переменной. Преодоление этих ограничений в исследовании факторных зависимостей обеспечивают методы интеллектуального анализа данных (data mining). Обобщая множество представленных в литературе толкований этого понятия, можно дать его следующее определение: *интеллектуальный анализ данных* - это извлечение, количественная и содержательная оценка осмысленных «паттернов» - априорно неизвестных и непредвиденных структур исследуемого многомерного объекта, а также моделей зависимостей характеризующих его признаков.

Метод «случайный лес» входит в арсенал методов data mining, поскольку позволяет выявить скрытые, не устанавливаемые априорно закономерности взаимосвязей и структуры в многомерном признаковом пространстве, характеризующем исследуемую совокупность. В аспекте поставленной задачи метод «случайный лес» позволяет получить следующие результаты:

- сформировать классификацию субъектов Российской Федерации по уровню неформальной занятости, обусловленную влиянием макроэкономических факторов регионального развития, которые находятся в сложной и в значительной степени латентной иерархической взаимосвязи;

- построить регрессию уровня неформальной занятости на региональном уровне по показателям, характеризующим макроэкономические параметры регионального развития, с учетом отмеченных особенностей их взаимосвязей.

Основные понятия и алгоритмы метода «случайный лес»

В широком плане можно определить, что метод «случайный лес» позволяет решать две основные задачи:

- классификации,

- регрессии.

Метод основан на построении большого числа (ансамбля) деревьев решений, каждое из которых строится по выборке, получаемой из исходной обучающей выборки с помощью бутстрепа (то есть выборки с возвращением) [10, с. 117-136].

Существует большое число публикаций, в которых изложена история развития метода «случайный лес» и отмечен вклад его основателей: создателя первого алгоритма для лесов случайных решений Тина Кам Хо [11], а также Лео Бреймана [12-14] и Адель Катлер [15] - авторов, сформировавших один из наиболее востребованных алгоритмов машинного обучения - Random Forest[7], заключающийся в использовании комитета (ансамбля) решающих деревьев.

Исходным теоретическим понятием и базовой структурной единицей метода «случайный лес» является «дерево решений».

Метод построения дерева решений применяется для совокупности единиц, каждая из которых характеризуется целевой (зависимой) перемен-

[7] Л. Брейман и А. Катлер зарегистрировали «Random Forest» в качестве товарного знака, который с 2019 г. принадлежит Minitab, Inc. URL: https://ru.qwe.wiki/wiki/Random_forest.

ной (y_i) и набором переменных-предикторов (x_i - факторных переменных). Последние определяют распределение единиц наблюдения по классам (группам) и влияние этого распределения на значения зависимой переменной.

При этом если целевая переменная дискретная (так называемая метка класса), то модель является деревом классификации, а если непрерывная, то деревом регрессии.

Дерево решений формируется на обучающем множестве, то есть деревья решений являются моделями, строящимися на основе «обучения с учителем». Оценка качества построения дерева по результатам обучения проводится на тестовой (контрольной) выборке.

Деревья решений представляют собой иерархические древовидные структуры (графы), полученные в результате применения к единицам совокупности алгоритма классификации, состоящего из решающих правил типа «если ..., то ...» для значений группирующих признаков этих единиц (предикторов).

Основные термины, используемые при описании алгоритма построения дерева решений, содержатся в таблице 3.

Математическая постановка построения дерева решений и оценки предсказанных значений зависимой переменной на его основе состоит в следующем [10]:

1. Деревом решений называется дерево, с каждой вершиной t которого связано некоторое подмножество $X_t \subset X$; с корневой вершиной связывается все пространство единиц X.

2. Задается некоторая функция (*решающее правило*) $f_t: X \to \{0, 1, \ldots, k_t - 1\}$, определяющая разбиение множества X на k непересекающихся подмножеств; (здесь $k_t > 2$ - количество потомков вершины t). С терминальными вершинами не связывается никакая функция.

В случае решения задачи классификации вектор x относится к классу, являющемуся мажорантным (наиболее часто встречающимся) в подвыборке D_t, соответствующей данной терминальной вершине; в случае регрессии оценка условного математического ожидания отклика представляет собой среднее значение отклика в этой подвыборке (в этом случае дерево решений часто называют деревом регрессий).

Критериями качества построения дерева являются следующие критерии «на минимум»:

- Для дерева регрессии:
  - сумма квадратов остатков (*RSS*), определяемая по формуле:

$$RSS = \sum_i (y_i - \bar{y})^2,$$

где RSS - сумма квадратов остатков (отклонений от средней); i - номер единицы в узле дерева регрессии; y_i - фактические значения результативной переменной в узле; $\bar{y}$ - среднее значение переменной y_i в рассматриваемом узле.

- Для дерева классификации:
  - частота ошибок классификации (E), то есть доля наблюдений в тестовой выборке, которые не принадлежат к наиболее распространенному классу по значениям предикторов, определивших этот класс в обучающей выборке;
  - индекс Джини (загрязнение Джини) (L_G) - вероятность неверной маркировки, то есть отнесения единицы к определенному классу.

    Индекс Джини можно интерпретировать как меру общей дисперсии во всех классах:

$$L_G = 1 - \sum_k (p_{km})^2.$$

    К примеру, если в результате расщепления корневого узла, включающего 6 единиц, образовалось два узла, в которые вошли, соответственно, 2 и 4 единицы (см. рис. 2), то значение загрязнения Джини равно:

$$L_G = 1 - [(2/6)^2 + (4/6)^2] = 0{,}444.$$

    По мере продвижения вниз по дереву загрязнение Джини должно снижаться, приняв нулевое значение в термальных узлах (листьях).
  - коэффициент перекрестной энтропии (D):

$$D = -\sum_k (p_{km}) \log (p_{km}).$$

    Коэффициент перекрестной энтропии принимает значение, близкое к нулю, когда все p_{km} близки к нулю или единице. Следовательно, как и индекс Джини, коэффициент перекрестной энтропии будет низким в случае «чистого» m-го узла [16].

Таблица 3

Основные термины, используемые в методе построения дерева решений

Объект	Наблюдение, единица совокупности, образец
Атрибут	Признак, независимая переменная
Целевая переменная	Зависимая переменная: количественная (отклик) или качественная (метка класса)
Узел	Внутренний узел дерева, узел проверки
Корневой узел (корневая вершина)	Начальный узел дерева решений
Терминальный узел (терминальная вершина), лист	Конечный узел дерева, узел решения - узел без исходящих связей
Решающее правило	Условие в узле, проверка
Критерии качества построения дерева	Для деревьев регрессий: • *RSS* (residual sum of squares) - сумма квадратов остатков Для деревьев классификации: • частота ошибок классификации: доля обучающих наблюдений в соответствующей области, которые не принадлежат к наиболее распространенному классу • индекс Джини - мера общей дисперсии во всех классах • энтропия

«Случайный лес» - модель, состоящая из множества деревьев решений, при построении которой учитываются следующие правила построения:

1. Используется случайная (повторная) выборка единиц из исходной совокупности при построении деревьев.

2. Дерево строится до исчерпания выборки (без «обрезки»).

3. При разделении узлов выбираются случайные наборы параметров.

«Случайный лес» - это *«ансамбль» деревьев*, то есть их комбинация, позволяющая получить предсказанные значения зависимой переменной.

Как было указано выше, если на основе метода «случайный лес» решается задача классификации *для качественной переменной*, то решение принимается на основе «*большинства голосов*»: окончательное предсказание представляет собой наиболее часто встречающийся класс среди всех полученных предсказаний.

Если решается задача регрессии для *количественной переменной отклика*, то предсказанное значение определяется как среднее арифметическое (простое или взвешенное) из всех значений отклика в термальных узлах, содержащих заданные значения предикторов, для которых определяется прогноз зависимой переменной (см. рис. 2).

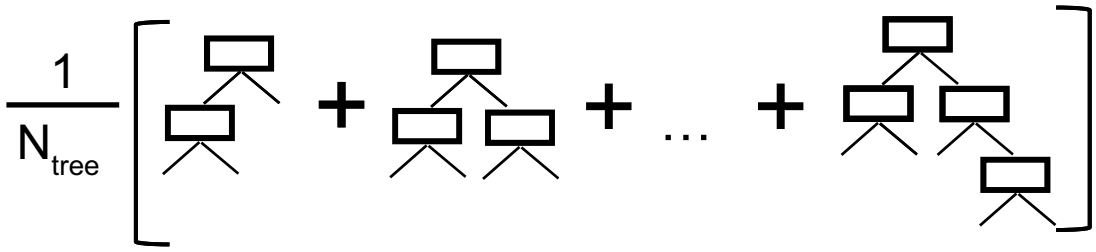


Рис. 2. **Случайный лес как «ансамбль» деревьев регрессии**

Источник: URL: http://www.machinelearning.ru/wiki/images/c/cc/PZAD2016_09_rf.pdf.

Применение изложенных алгоритмов и критериев оценки результатов для решения задач классификации регионов по уровню неформальной занятости и регрессии этого показателя по макроэкономическим параметрам регионального развития представлено ниже.

Результаты классификации и регрессии методом «случайный лес» в исследовании влияния макроэкономических показателей регионального развития на уровень неформальной занятости

Применение алгоритмов метода «случайный лес» для решения прикладных статистических задач возможно с использованием пакета «randomForest» системы R, а также модуля «Random Forest» в подсистеме «Data Mining» пакета STATISTICA.

Начальным этапом исследования является группировка субъектов по показателю «Удельный вес неформально занятых лиц в общей численности занятых, в процентах».

В соответствии с гистограммой, представленной на рис. 3, субъекты Российской Федерации априорно были разделены на три группы:

1-я группа - до 20% (33 региона);

2-я группа - 20-30% (35 регионов);

3-я группа - свыше 30% (17 регионов).

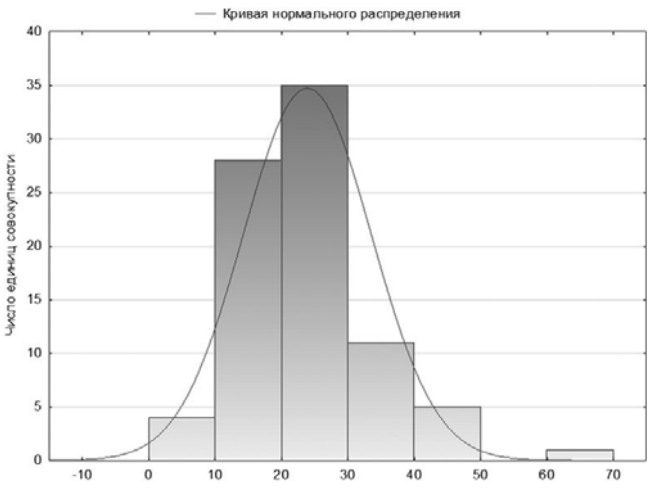


Рис. 3. **Гистограмма распределения субъектов Российской Федерации по показателю «Удельный вес неформально занятых лиц в общей численности занятых, в процентах»**

Согласно вышеизложенной постановке задачи классификации для формирования «ансамбля» задавалось построение 100 деревьев классификаций на обучающей совокупности, которая составляла 70% от исходной совокупности.

На рис. 4 представлена диаграмма сходимости результатов классификаций, полученных по обучающей и тестовой выборкам, согласно которой минимизация расхождений достигается при построении 80 деревьев.

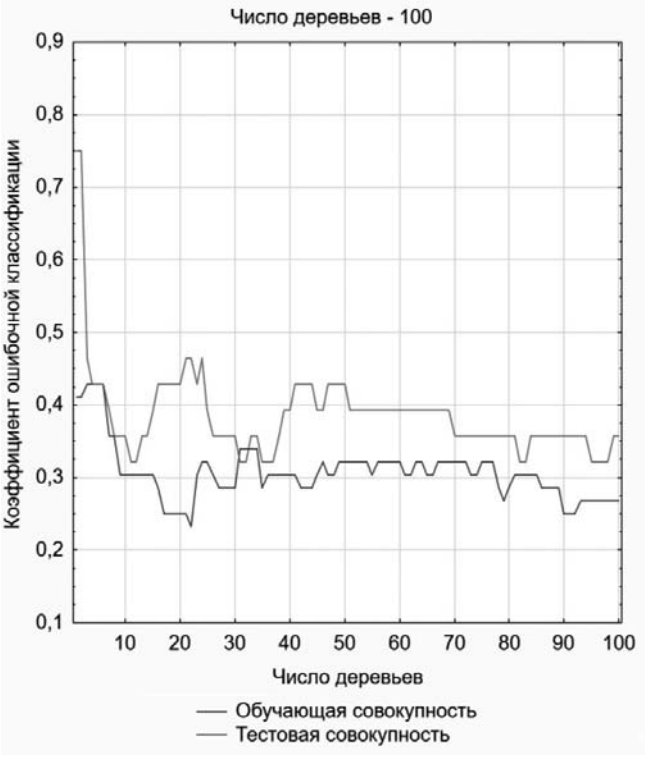
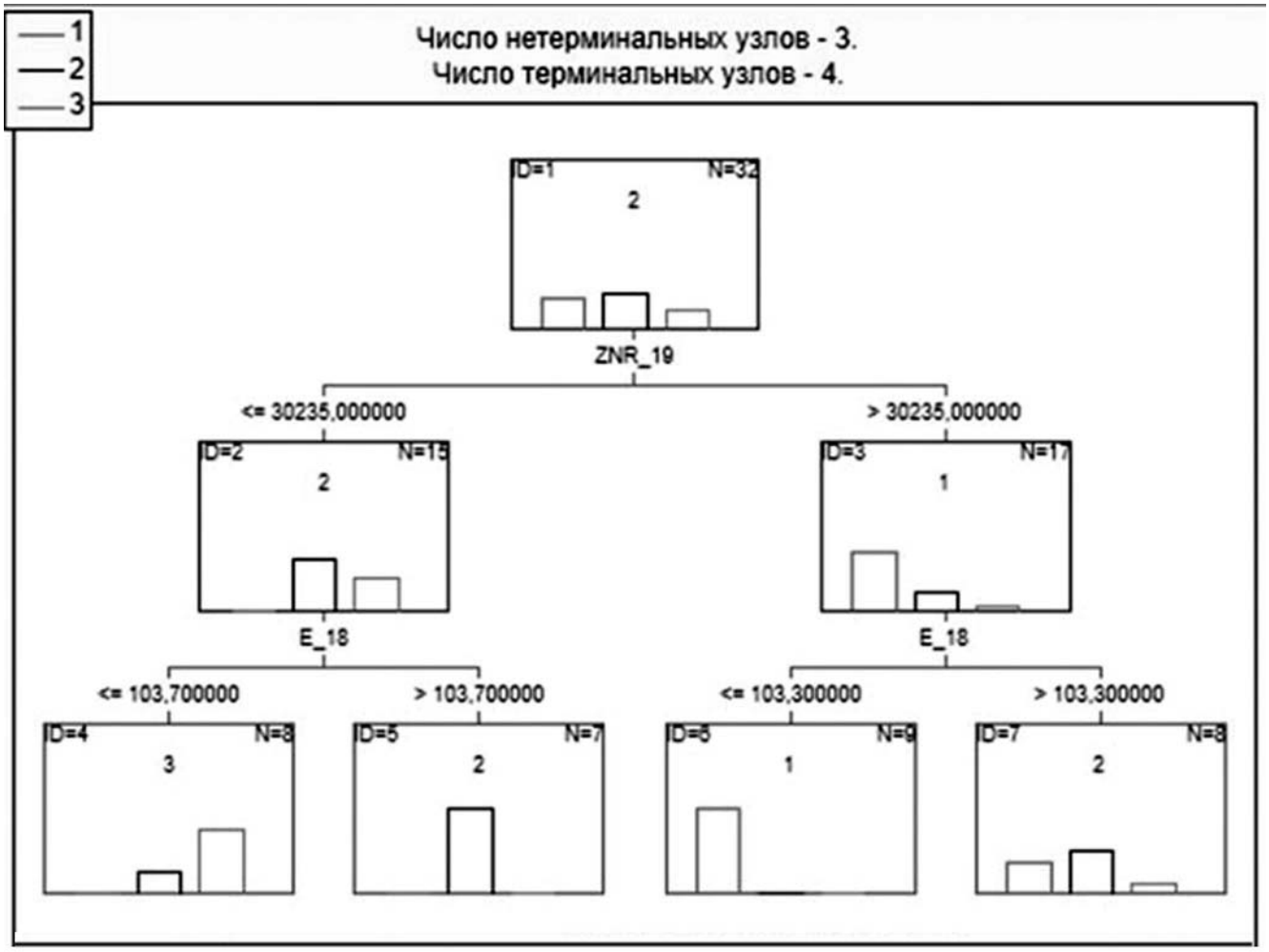


Рис. 4. **Оценки ошибки классификации в зависимости от числа деревьев, объединяемых в «ансамбль»**

На рис. 5-7 представлены отдельные деревья классификации, построенные по тестовой совокупности и дающие представление о иерархической взаимосвязи исследуемых факторных переменных.

Число узлов в каждом дереве определялось условием достижения минимума приведенного выше критерия «чистоты» Джини. В начальном узле представлена диаграмма исходного распределения регионов на группы по значениям показателя «Удельный вес неформально занятых лиц в общей численности занятых, в процентах», и в изображении каждого нового узла представлена диаграмма распределения регионов в зависимости от значений факторных показателей, определивших формирование узла.

Граф дерева классификации, изображенный на рис. 5, свидетельствует о том, что в 2019 г. критериальный уровень среднемесячного трудового дохода (ZNR), определяющий разделение регионов на две подгруппы первого уровня по значению показателя неформальной занятости, составлял 30235 рублей. В регионах с меньшим значением этого факторного показателя удельный вес неформально занятых лиц в общей численности занятых превышает 20%. Соответственно,

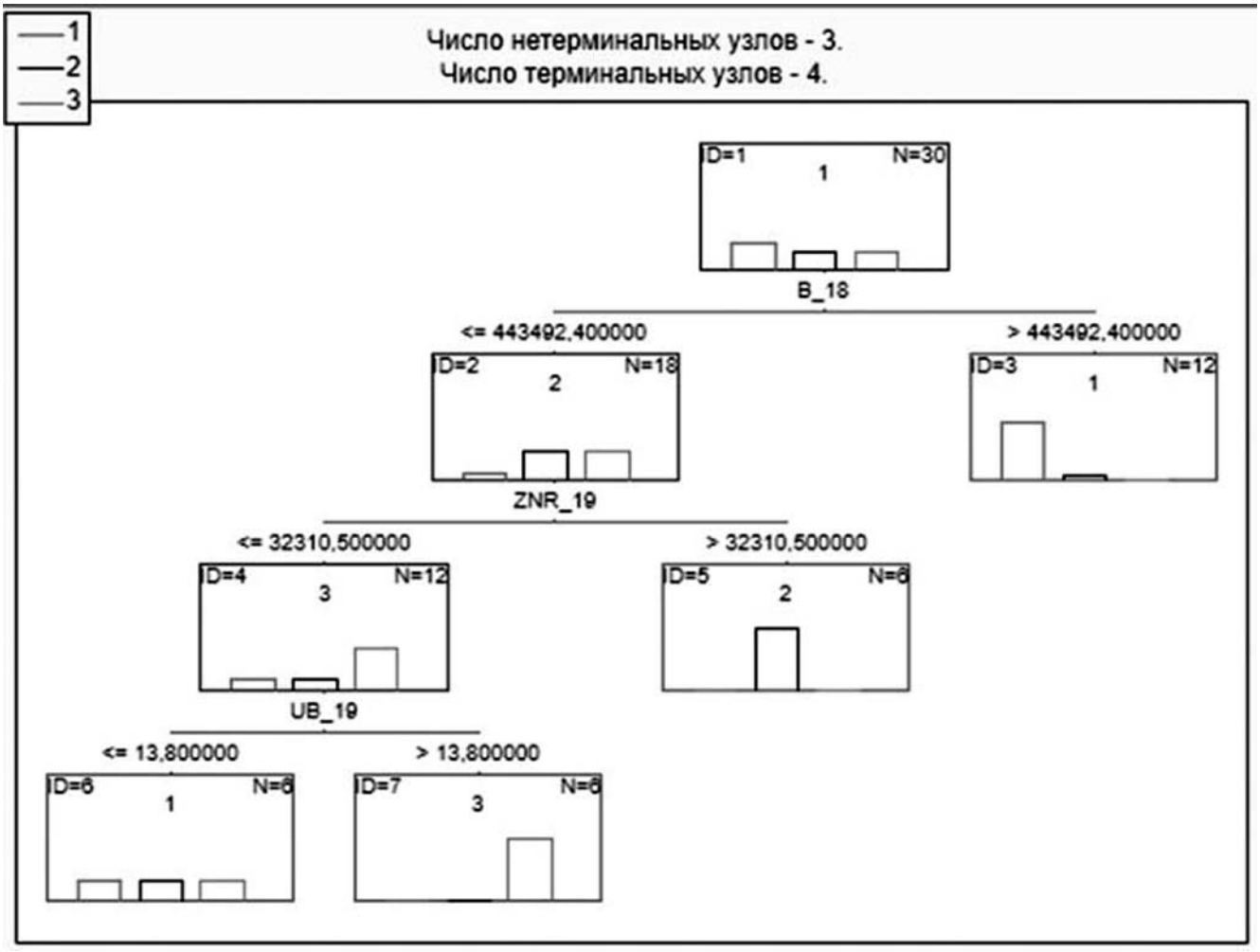


Условные обозначения: ZNR_19 - среднемесячный трудовой доход (заработная плата наемных работников в формальном и неформальном секторах экономики), рублей, 2019 г.; E_18 - индекс производительности труда, в процентах, 2018 г.

Рис. 5. **Дерево классификации, представляющее группировку регионов по значениям показателя «Удельный вес неформально занятых лиц в общей численности занятых, в процентах» в зависимости от значений и иерархических связей факторных переменных, 2019 г.**

в регионах, где среднемесячный трудовой доход больше указанного критериального значения, уровень неформальной занятости составил менее 20%. При этом для обеих выделенных подгрупп значимым является показатель «Индекс производительности труда, в процентах» с годовым лагом запаздывающего влияния. В первой подгруппе годовой темп роста производительности 103,7% разделяет на втором уровне регионы на две подгруппы: с долей неформальной занятости более 30% там, где темп роста производительности оказался менее 103,7%, и, соответственно, с долей неформальной занятости от 20 до 30% в тех случаях, когда темп роста производительности превысил 103,7%.

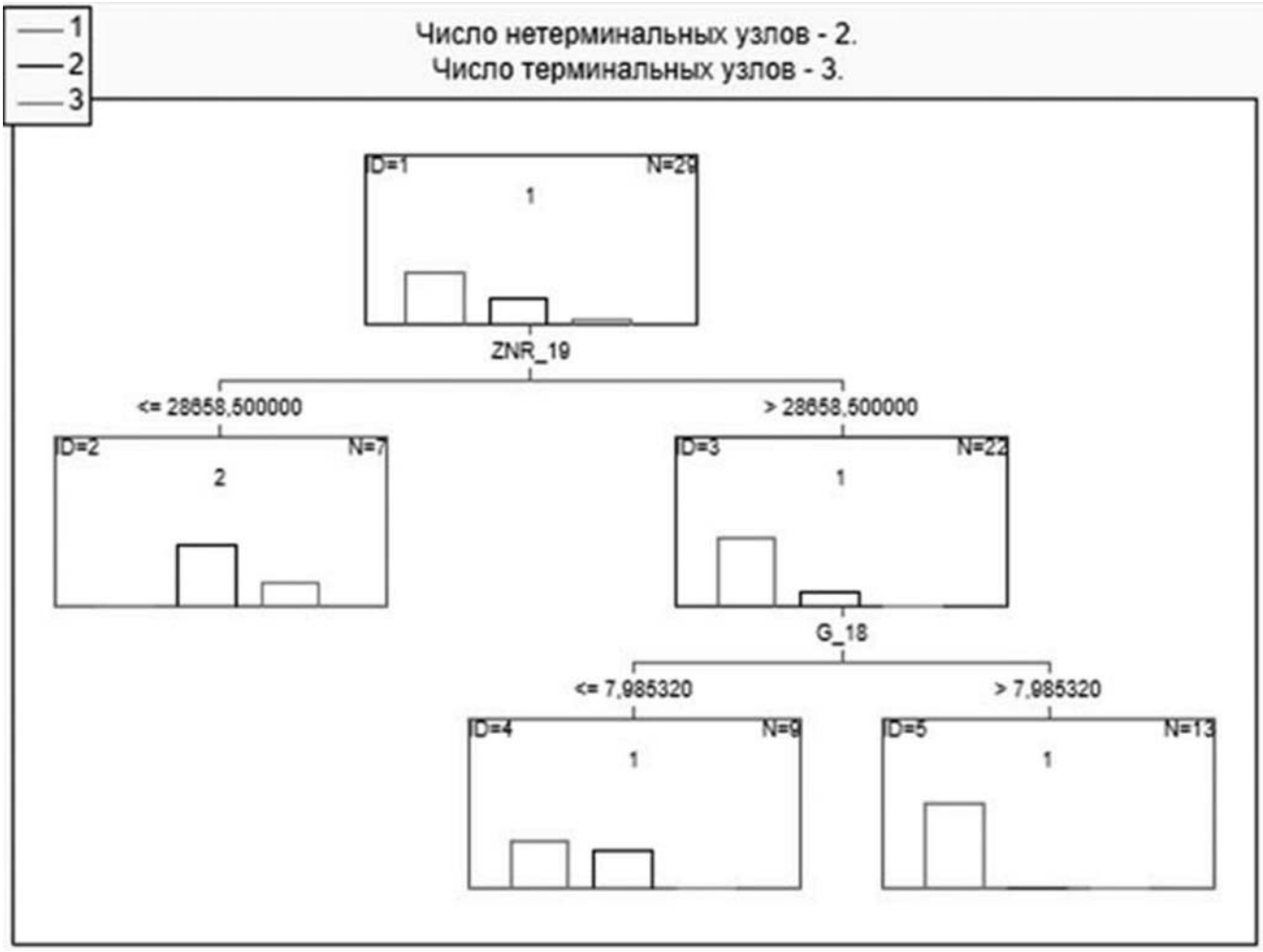
Таким образом, в регионах с относительно низким среднемесячным трудовым доходом (среднемесячной заработной платы в формальном и неформальном секторах экономики) темпы роста производительности труда имеют обратную статистическую связь с уровнем неформальной занятости. Но в регионах с относительно более высоким уровнем среднемесячного трудового дохода проявилась прямая статистическая зависимость уровня неформальной занятости и динамики производительности труда: при значении темпа роста производительности труда менее 103,7% уровень неформальной занятости составил менее 20%; превышение темпа роста производительности над этим критериальным значением обусловливает (хотя и с невысокой степенью статистической значимости) повышение уровня неформальной занятости до уровня 20% и выше.



Условные обозначения: B-18 - валовой региональный продукт на душу населения, рублей, 2018 г.; ZNR_19 - среднемесячный трудовой доход (заработная плата наемных работников в формальном и неформальном секторах экономики), рублей, 2019 г.; UB_19 - уровень бедности (доля населения со среднедушевыми денежными доходами ниже величины прожиточного минимума), в процентах, 2019 г.

Рис. 6. **Дерево классификации, представляющее группировку регионов по значениям показателя «Удельный вес неформально занятых лиц в общей численности занятых, в процентах» в зависимости от значений и иерархических связей факторных переменных, 2019 г.**

На рис. 6 представлен результат, соответствующий дереву классификации и указывающий на то, что если в регионе уровень бедности превышает 13,8% при условии, что среднемесячный трудовой доход не достигает значения 32310,5 рублей, то уровень неформальной занятости составляет 30% и выше. В регионах со среднемесячным трудовым доходом свыше 32310,5 рублей типичный уровень неформальной занятости - 20-30%. При этом необходимо отметить, что указанное деление характерно для регионов с объемом ВРП на душу населения менее 443492,4 рублей. В регионах с бо́льшим значением данного показателя (более высоким темпом экономического развития) уровень как среднемесячного трудового дохода, так и бедности не оказывают статистически значимого влияния на процент неформально занятых лиц; их удельный вес составляет менее 20% и определяется другими (внутрирегиональными) факторами.



Условные обозначения: ZNR_19 - среднемесячный трудовой доход (заработная плата наемных работников в формальном и неформальном секторах экономики), рублей, 2019 г.; G_18 - доля работников микропредприятий в ССЧ, в процентах, 2018 г.

Рис. 7. **Дерево классификации, представляющее группировку регионов по значениям показателя «Удельный вес неформально занятых лиц в общей численности занятых, в процентах» в зависимости от значений и иерархических связей факторных переменных, 2019 г.**

Граф, представленный на рис. 7, свидетельствует о том, что при превышении значения среднемесячного трудового дохода в регионах 28628,5 рублей значимым факторным показателем снижения уровня неформальной занятости (от 20% и ниже) является «Доля работников микропредприятий в ССЧ». Именно в этой группе регионов (со среднемесячным трудовым доходом в экономике свыше 28628,5 рублей) по результатам проведенных расчетов проявляется эффект легализации труда самозанятых лиц: оформление ими индивидуальных и микропредприятий.

На рис. 8 представлена диаграмма распределения предикторов (факторных показателей) по значению критерия «важность», из которой следует, что наибольшее влияние на итоговую классификацию субъектов Российской Федерации по уровню неформальной занятости оказывают следующие факторные показатели: валовой региональный продукт на душу населения, рублей; среднемесячная заработная плата в организациях полного круга, рублей; среднемесячный трудовой доход (заработная плата наемных работников в формальном и неформальном секторах экономики), рублей; доля вида деятельности «Сельское, лесное хозяйство, охота, рыболовство и рыбоводство» в отраслевой структуре ВРП; доля вида деятельности «Образование» в отраслевой структуре ВРП; фактическое конечное потребление домашних хозяйств в расчете на душу населения на территории субъектов Российской Федерации (в текущих рыночных ценах), рублей.

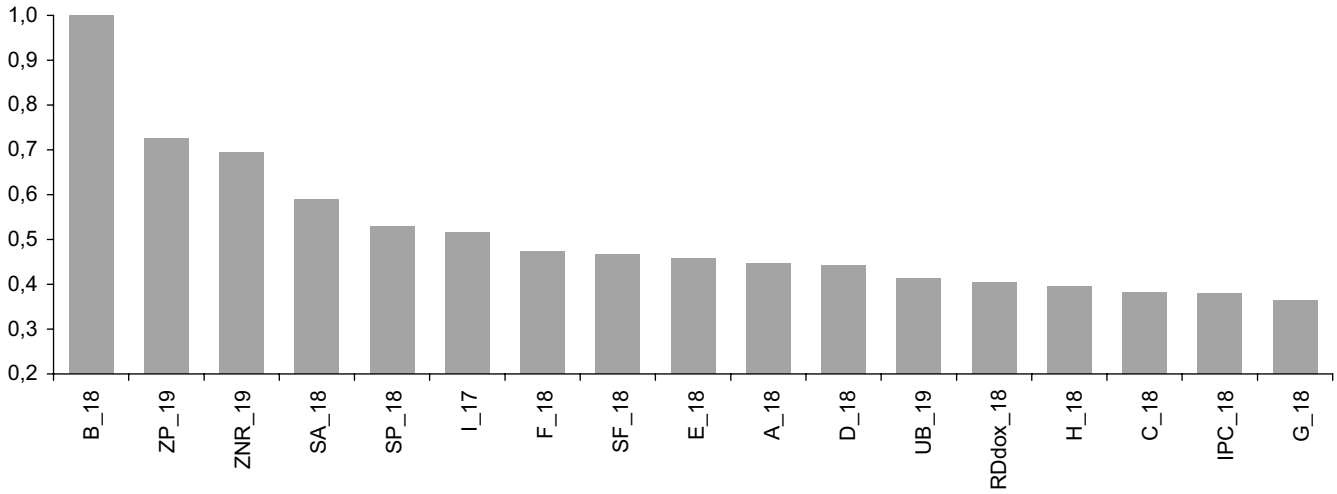


Рис. 8. **Диаграмма значимости предикторов классификации в масштабе 0-1**

Картограммы (рис. 9 и 10) представляют исходную и итоговую группировки субъектов Российской Федерации по уровню неформальной занятости. Итоговая классификация выделяет группы регионов, однородные по уровню неформальной занятости, каждая из которых обусловлена специфическим набором факторных переменных, особенностями их иерархической взаимозависимости и влияния на результативную переменную.

Алгоритмы метода «случайный лес» позволяют строить регрессию зависимой величины от переменных-предикторов с учетом их иерархических взаимосвязей. Можно сформировать логическую цепочку последовательности методов многофакторного регрессионного моделирования зависимых переменных на региональном уровне:

1 - Панельная регрессия - единый набор факторных переменных для всех регионов;

Рис. 9. **Исходная группировка субъектов Российской Федерации по показателю «Удельный вес неформально занятых лиц в общей численности занятых, в процентах»**

Примечание (для рис. 9 и 10): выделено три группы в соответствии с распределением результативной переменной: 1-я группа - до 20%; 2-я группа - от 20 до 30%; 3-я группа - свыше 30%.

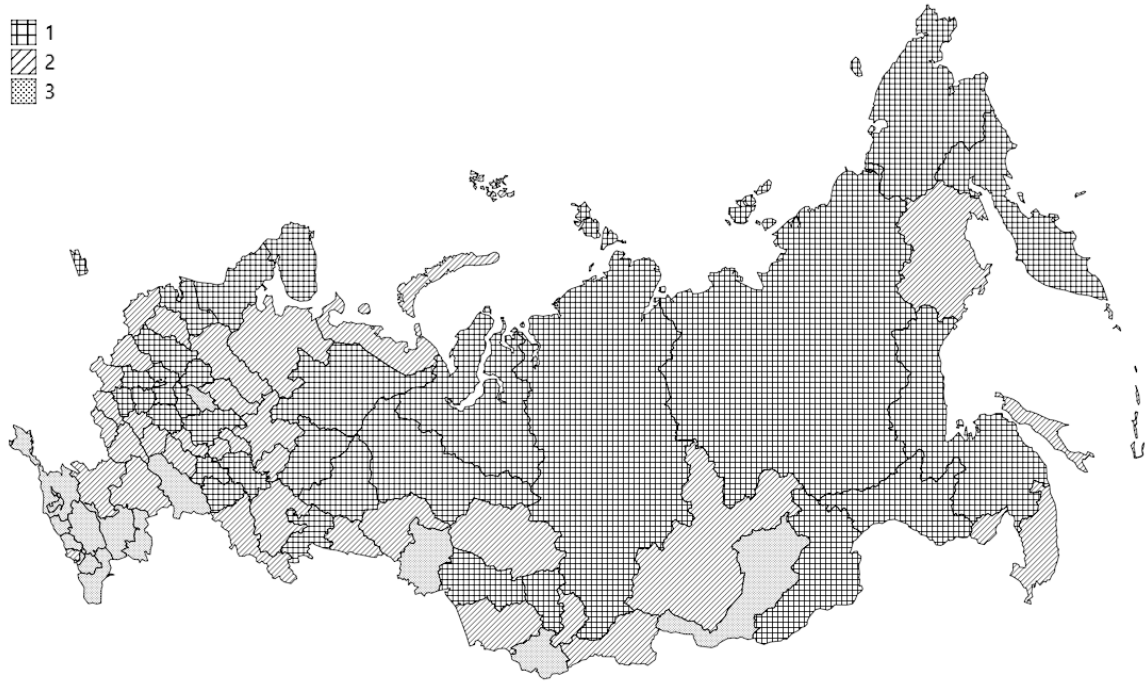


Рис. 10. **Итоговая классификация субъектов Российской Федерации по уровню неформальной занятости, определяемому иерархической взаимозависимостью макроэкономических показателей регионального развития**

специфика отдельных регионов отражается через перераспределение общего смещения модели через так называемые «фиксированные эффекты».

2 - Регрессионные модели по выделенным на основе анализа множества факторных переменных региональным кластерам.

В этом случае наборы факторных переменных и параметры моделей отличаются по региональным кластерам, что позволяет учесть групповую специфику моделируемых закономерностей факторного влияния на результативную величину.

3 - Регрессии на факторные переменные на основе построения «ансамбля» деревьев решений методом «случайный лес». В этом случае специфика регионов учитывается в наибольшей степени: набор факторных переменных, их значения, характер взаимозависимости являются особыми для отдельных «узких» гомогенных подгрупп регионов, которые выделяются на основе критерия минимальности остаточной дисперсии.

Как следует из рис. 11, сходимость регрессий обучающей и тестовой совокупностей при построении «ансамблей» деревьев регрессий обеспечивается на 53-55 деревьях из рассматриваемых в примере 100 деревьев.

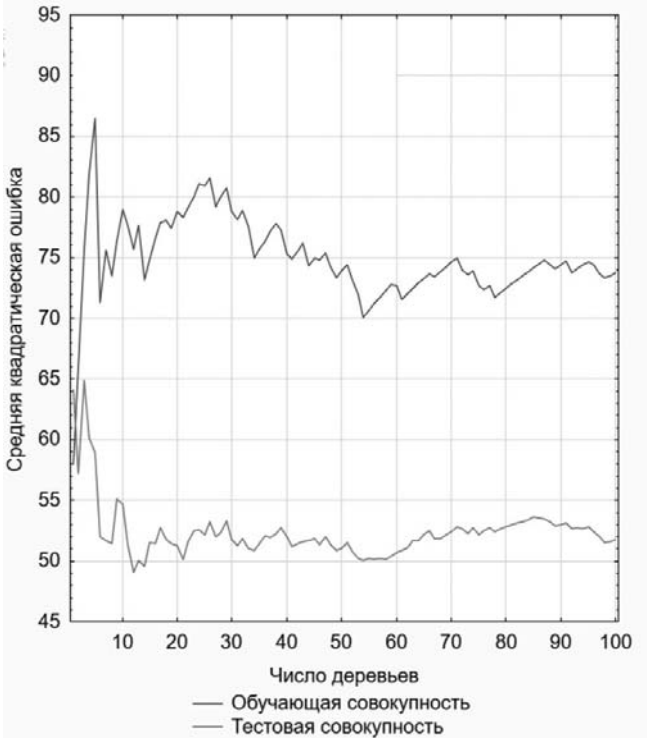


Рис. 11. **Оценки ошибки регрессии в зависимости от числа деревьев, объединяемых в «ансамбль»**

Ранжированное распределение факторных переменных по уровню значимости для построения регрессии методом «случайный лес» представлено на рис. 12.

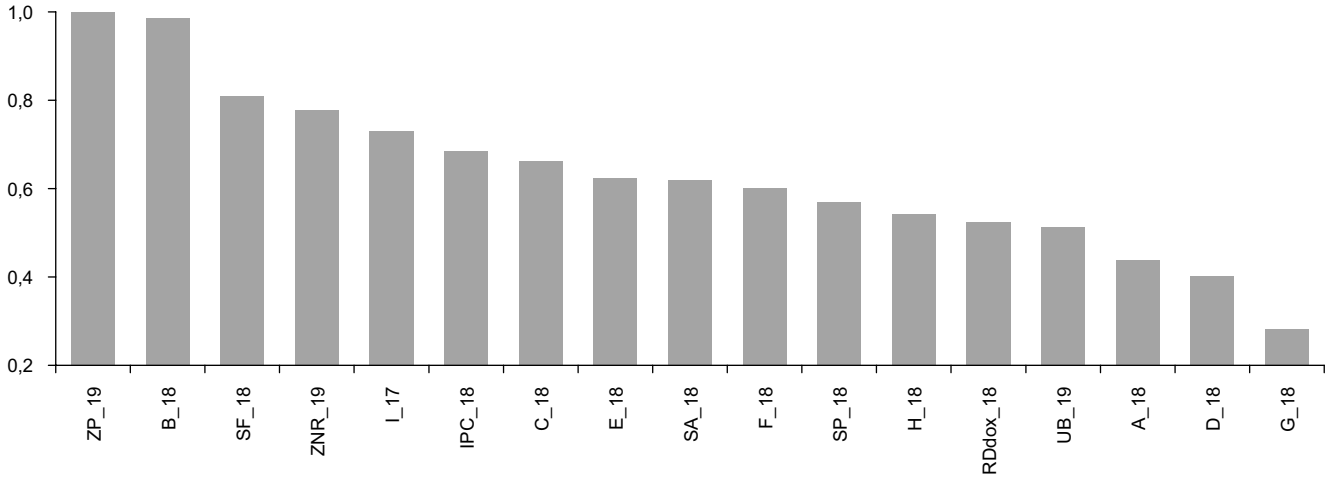
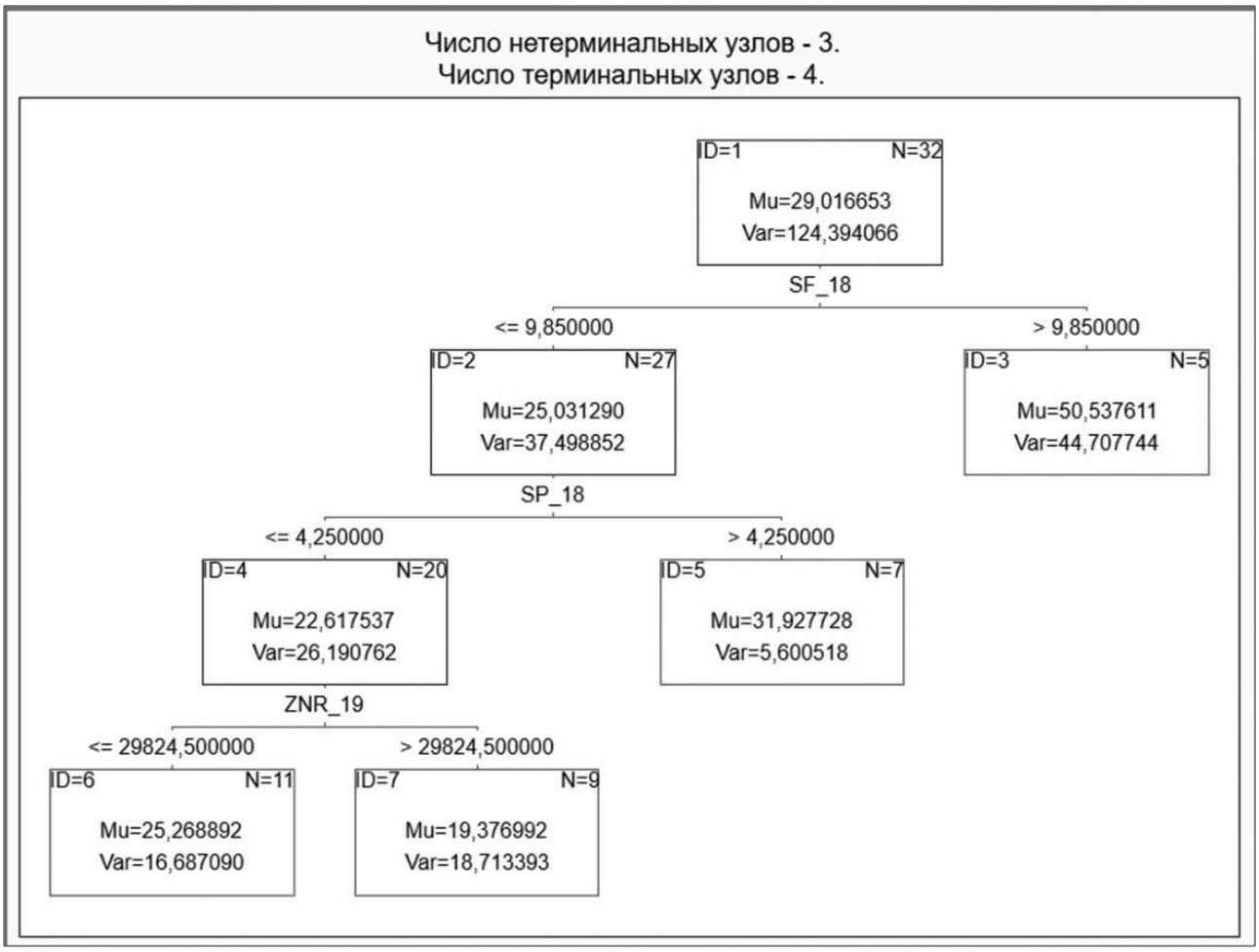


Рис. 12. **Диаграмма значимости предикторов регрессии в масштабе 0-1**



Условные обозначения: SF_18 - доля вида деятельности «Строительство» в отраслевой структуре ВРП, в процентах, 2018 г.; SP_18 - доля вида деятельности «Образование» в отраслевой структуре ВРП, в процентах, 2018 г.; ZNR_19 - среднемесячный трудовой доход (заработная плата наемных работников в формальном и неформальном секторах экономики), рублей, 2019 г.

Рис. 13. **Дерево регрессии показателя «Удельный вес неформально занятых лиц в общей численности занятых, в процентах» по регионам России, 2019 г.**

Примечание: в узлах представлены средние значения зависимой переменной (Mu), а также значения показателя дисперсии (Var).

Из сравнения рис. 8 и 12 следует вывод, что в рассматриваемом примере для построения качественной регрессии показателя «Удельный вес неформально занятых лиц в общей численности занятых, в процентах» более значимыми, чем для построения качественной классификации по значениям этого показателя, являются такие факторные переменные, как «Доля вида деятельности "Строительство" в отраслевой структуре ВРП, в процентах»; «Индекс потребительских цен, в процентах», «Доля валового регионального продукта субъекта Российской Федерации в валовом региональном продукте РФ, в процентах». Менее значимыми при построении регрессии оказались факторные переменные «Уровень бедности, в процентах», «Доля работников малых предприятий в общей численности ССЧ, в процентах».

На рис. 13 представлен вариант дерева регрессии, демонстрирующий прямую связь уровня неформальной занятости в регионах с показателями удельного веса отраслей «Строительство» и «Образование» в структуре ВРП, что указывает на то, что экономические единицы этих отраслей являются «центрами притяжения» неформальной занятости, которая создает конкуренцию формально организованному труду в этих отраслях. Критериальными значениями этих факторных показателей для «расслоения» регионов по уровню неформальной занятости являются, соответственно, значения 9,85 и 4,25%. В регионах с низкими относительно этих критериев значениями удельного веса указанных отраслей в объеме ВРП значимым факторным показателем, формирующим регрессию уровня неформальной занятости, становится показатель среднемесячного трудового дохода (заработная плата наемных работников в формальном и неформальном секторах экономики).

Сравнение расчетных значений показателя «Удельный вес неформально занятых лиц в общей численности занятых, в процентах» за 2019 г., полученных по результатам построения «ансамбля» деревьев регрессии, с соответствующими фактическими значениями представлено на рис. 14.

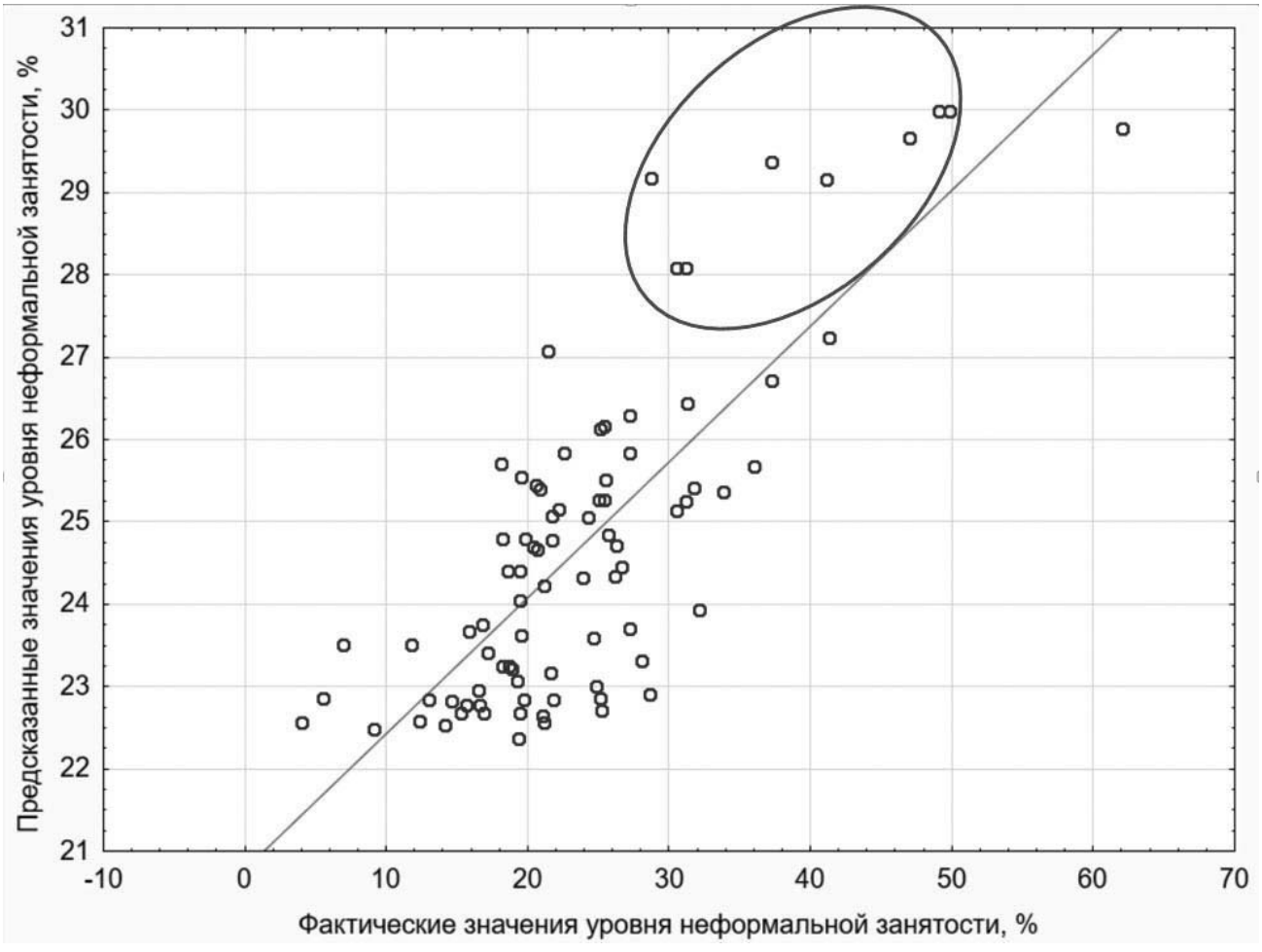


Рис. 14. **Сопоставление расчетных (предсказанных) значений показателя уровня неформальной занятости в регионах России, полученных по результатам построения «ансамбля» деревьев регрессии, с соответствующими фактическими значениями, 2019 г.**

Как следует из данных рис. 14, построенная регрессия является адекватной для диапазона, не превышающего уровень 30% удельного веса неформально занятых лиц в общей численности занятых (14% субъектов Российской Федерации). Вместе с тем для регионов, имеющих уровень неформальной занятости свыше 30%, необходимо уточнить спецификацию регрессионной модели неформальной занятости.

Качество полученной методом «случайный лес» регрессионной модели подтверждается соответствием распределения остатков нормальному закону распределения при 10%-м уровне значимости (см. рис. 15).

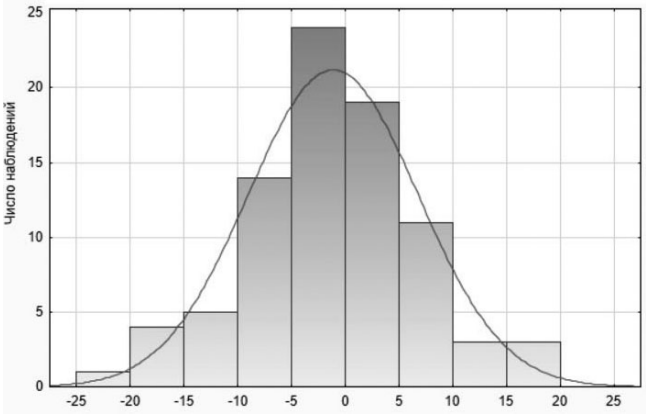


Рис. 15. **Оценка соответствия остатков регрессии показателя «Удельный вес неформально занятых лиц в общей численности занятых, в процентах», построенной по регионам России методом «случайный лес», теоретически нормальному закону распределения**

Заключение

Представленные результаты свидетельствуют о получении адекватной классификации регионов России по уровню неформальной занятости, а также статистически надежной регрессии этого показателя от макроэкономических показателей регионального развития. Более высокое качество этих статистических решений по сравнению с результатами применения традиционных методов обеспечивается учетом изначально скрытых от исследователя, но выявляемых методом «случайный лес» иерархических взаимозависимостей факторных показателей и их многоуровневых связей с зависимой переменной. Данные результаты являются теоретически значимыми для дальнейших исследований региональных факторов, влияющих на неформальную занятость, а также имеют практическую ценность, поскольку реализованы на данных официальной региональной статистики и предлагают новые возможности информационного обеспечения принятия решений в сфере государственной политики формализации труда.

Литература

1. **Санги А., Фрейхе-Родригес С., Пошарац А.** Проблема неформальной занятости в России. Причины и варианты решения. Группа Всемирного банка, 2019. URL: http://documents1.worldbank.org/curated/en/835091559937396870/pdf/Stemming-Russia-s-Informality-Unearthing-Causes-and-Developing-Solutions.pdf.

2. **Cohen J.E.** Macroeconomic and Microeconomic Determinants of Informal Employment: The Case of Clothing Traders in Johannesburg, South Africa. PhD Diss. (Econ.), Amherst, University of Massachusetts, 2012. URL: https://scholarworks.umass.edu/dissertations/AAI3545913.

3. **Huang G., Xue D., Wang B.** Integrating Theories on Informal Economies: An Examination of Causes of Urban Informal Economies in China // Sustainability. 2020. Vol. 12. Iss. 7. P. 2738. doi: https://doi.org/10.3390/su12072738.

4. **Maddah M., Sobhani B.** The Effective Factors on Informal Economy in Developing Countries (Panel Data Model) // International Journal of Regional Development. 2014. Vol. 1. No. 1. P. 12-25. doi: https://doi.org/10.5296/ijrd.v1i1.6437.

5. **Kucera D., Roncolato L.** Informal Employment: Two Contested Policy Issues // International Labour Review. 2008. Vol. 147. Iss. 4. P. 321-348. doi: https://doi.org/10.1111/j.1564-913X.2008.00039.x.

6. **Lavoie M., Stockhammer E.** Wage-led Growth: Concept, Theories and Policies // M. Lavoie, E. Stockhammer (eds). Wage-led Growth. Advances in Labour Studies. London: Palgrave Macmillan, 2013. URL: https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/@publ/documents/genericdocument/wcms_234602.pdf.

7. ILO Department of Statistics. Statistical Update on Employment in the Informal Economy. ILO, June 2011. URL: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/presentation/wcms_157467.pdf.

8. OECD/ILO. Tackling Vulnerability in the Informal Economy. Paris: OECD Publ., 2019. doi: https://doi.org/10.1787/939b7bcd-en.

9. **Benjamin N., Mbaye A.A.** The Informal Sector in Francophone Africa: Firm Size, Productivity, and Institutions. Washington, DC: World Bank, 2012. doi: https://doi.org/10.1596/978-0-8213-9537-0.

10. **Чистяков С.П.** Случайные леса: обзор // Труды Карельского научного центра РАН. 2013. № 1. С. 117-136. URL: http://resources.krc.karelia.ru/transactions/doc/trudy2013/trudy_2013_1_117-136.pdf.

11. **Ho T.K.** The Random Subspace Method for Constructing Decision Forests // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998. Vol. 20. No. 8. P. 832-844. doi: https://doi.org/10.1109/34.709601.

12. **Breiman L.** et al. Classification and Regression Trees. Wadsworth, New York: Chapman and Hall, 1984.

13. **Breiman L.** Bagging Predictors // Machine Learning. 1996. Vol. 24. Iss. 2. P. 123-140. doi: https://doi.org/10.1023/A:1018054314350.

14. **Breiman L.** Random Forests // Machine Learning. 2001. Vol. 45. Iss. 1. P. 5-32. doi: https://doi.org/10.1023/A:1010933404324.

15. **Cutler A., Cutler R.D., Stevens J.R.** Random Forests // C. Zhang, Y. Ma (eds). Ensemble Machine Learning: Methods and Applications. Boston, MA: Springer, 2011. P. 157-175. doi: https://doi.org/10.1007/978-1-4419-9326-7_5.

16. **Михальченко Г.Е., Михальченко А.Г.** Дискретная математика: учеб. пособие / Сиб. федер. ун-т, Ин-т космич. и информ. технологий, Ин-т математики и фундамент. информатики. Красноярск: СФУ, 2011. 127 с.

Информация об авторах

Зарова Елена Викторовна - д-р экон. наук, профессор, заместитель руководителя проектного офиса, ГБУ «Аналитический центр» Правительства Москвы; профессор кафедры статистики, РЭУ им. Г.В. Плеханова. 119019, г. Москва, ул. Новый Арбат, д. 11, стр. 1; 117997, Москва, Стремянный пер., д. 36. E-mail: ZarovaEV@develop.mos.ru. ORCID: https://orcid.org/0000-0003-0375-2534.

Дубравская Эльвира Ивановна - главный эксперт, ГБУ «Аналитический центр» Правительства Москвы. 119019, г. Москва, ул. Новый Арбат, д. 11, стр. 1. E-mail: elvira.dubravskaya@yandex.ru. ORCID: https://orcid.org/0000-0003-3029-8111.

References

1. **Sanghi A., Freije-Rodriguez S., Posarac A.** *Stemming Russia's Informality: Unearthing Causes and Developing Solutions*. The World Bank Group; 2019. (In Russ.) Available from: http://documents1.worldbank.org/curated/en/835091559937396870/pdf/Stemming-Russia-s-Informality-Unearthing-Causes-and-Developing-Solutions.pdf.

2. **Cohen J.E.** *Macroeconomic and Microeconomic Determinants of Informal Employment: The Case of Clothing Traders in Johannesburg, South Africa*. PhD Diss. (Econ.), Amherst: University of Massachusetts; 2012. Available from: https://scholarworks.umass.edu/dissertations/AAI3545913.

3. **Huang G., Xue D., Wang B.** Integrating Theories on Informal Economies: An Examination of Causes of Urban Informal Economies in China. *Sustainability*. 2020;12(7):2738. Available from: https://doi.org/10.3390/su12072738.

4. **Maddah M., Sobhani B.** The Effective Factors on Informal Economy in Developing Countries (Panel Data Model). *International Journal of Regional Development*. 2014;1(1):12-25. Available from: https://doi.org/10.5296/ijrd.v1i1.6437.

5. **Kucera D., Roncolato L.** Informal Employment: Two Contested Policy Issues. *International Labour Review*. 2008;147(4):321-348. Available from: https://doi.org/10.1111/j.1564-913X.2008.00039.x.

6. **Lavoie M., Stockhammer E.** Wage-led Growth: Concept, Theories and Policies. In: M. Lavoie, E. Stockhammer (eds). *Wage-led Growth. Advances in Labour Studies*. London: Palgrave Macmillan; 2013. Available from: https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/@publ/documents/genericdocument/wcms_234602.pdf.

7. ILO Department of Statistics. *Statistical Update on Employment in the Informal Economy*. ILO; 2011. Available from: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/presentation/wcms_157467.pdf.

8. OECD/ILO. *Tackling Vulnerability in the Informal Economy*. Paris: OECD Publ.; 2019. Available from: https://doi.org/10.1787/939b7bcd-en.

9. **Benjamin N., Mbaye A.A.** *The Informal Sector in Francophone Africa: Firm Size, Productivity, and Institutions*. Washington, DC: World Bank; 2012. Available from: https://doi.org/10.1596/978-0-8213-9537-0.

10. **Chistiakov S.P.** Random Forests: An Overview. *Transactions of the Karelian Research Centre of the Russian Academy of Sciences*. 2013;(1):117-136. (In Russ.) Available from: http://resources.krc.karelia.ru/transactions/doc/trudy2013/trudy_2013_1_117-136.pdf.

11. **Ho T.K.** The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1998;20(8):832-844. Available from: https://doi.org/10.1109/34.709601.

12. **Breiman L**. et al. *Classification and Regression Trees*. Wadsworth, New York: Chapman and Hall; 1984.

13. **Breiman L.** Bagging Predictors. *Machine Learning*. 1996;24(2):123-140. Available from: https://doi.org/10.1023/A:1018054314350.

14. **Breiman L.** Random Forests. *Machine Learning*. 2001;45(1):5-32. Available from: https://doi.org/10.1023/A:1010933404324.

15. **Cutler A., Cutler R.D., Stevens J.R.** Random Forests. In: Zhang C., Ma Y. (eds). *Ensemble Machine Learning: Methods and Applications*. Boston, MA: Springer; 2011. P. 157-175. Available from: https://doi.org/10.1007/978-1-4419-9326-7_5.

16. **Mikhalchenko G.E., Mikhalchenko A.G.** *Discrete Mathematics: Textbook*. Siberian Federal University, SibFU Institute of Space and Information Technologies, Institute of Mathematics and Fundamental Informatics. Krasnoyarsk: Siberian Federal University Publ.; 2011. 127 p. (In Russ.)

About the authors

Elena V. Zarova - Dr. Sci. (Econ.), Professor, Deputy Head, Project Office, Analytical Center by Moscow City Government; Professor, Department of Statistics, Plekhanov Russian University of Economics. 11, New Arbat Ave., Bldg. 1, Moscow, 119019, Russia; 36 Stremyanny Lane, Moscow, 117997, Russia. E-mail: ZarovaEV@develop.mos.ru. ORCID: https://orcid.org/0000-0003-0375-2534.

Elvira I. Dubravskaya - Senior Analyst, Project Office of the Analytical Centre by Moscow City Government. 11, New Arbat Ave., Bldg. 1, Moscow, 119019, Russia. E-mail: elvira.dubravskaya@yandex.ru. ORCID: https://orcid.org/0000-0003-3029-8111.