Т.В. Кузьминова
(Москва)

МОДЕЛИРОВАНИЕ ДИНАМИКИ БЕЗРАБОТИЦЫ (по данным развития России 1996–2001 гг.)

Статья посвящена построению и анализу регрессионных и нейросетевых моделей динамики безработицы в России с учетом временных лагов. Предложен алгоритм поиска вида регрессионной зависимости по нескольким критериям, исключающим перебор всех возможных вариантов. Определены критерии построения нейросетевых моделей. Проведено сравнение прогнозных возможностей регрессионных и нейросетевых моделей.

Ключевые слова: регрессионная модель, нейросетевая модель, динамика, временной лаг, критерий качества модели.

Постановка задачи

Проблемы безработицы являются достаточно сложными во всем мире. Многие социологи и экономисты пытались решить их с различных точек зрения [1; 2; 3]. В России явление безработицы представляет особый интерес для исследования, так как не подчиняется многим тенденциям, характерным для других стран.

Как известно, российский рынок труда отличается наличием существенной по размерам теневой экономики [4, с. 14–24; 5, с. 232–248].

По данным Института макроэкономических исследований при Министерстве экономики Российской Федерации доходы «тени» составили в 1998 г. 1475 млрд. руб., в 1999 – 3100 млрд. руб. [4, с. 14–24].

Татьяна Владиславовна Кузьминова – кандидат технических наук, доцент, докторант кафедры прикладной социологии МГСУ.

Это составляет соответственно 54% и 65% годового валового внутреннего продукта страны.

Размеры теневой экономики влияют не только на саму безработицу, делая ее частично фиктивной, но и изменяют структуру безработицы, затрудняя прогнозирование развития ситуации на рынке труда.

Основным предположением проводимого исследования явилась возможность косвенно учесть влияние теневой экономики при рассмотрении безработицы как системного элемента социально-экономической системы. Известно, что системный подход заключается в рассмотрении всех элементов системы в их взаимосвязи друг с другом. Каждый элемент в системе обладает свойствами, отличными от тех, что характеризуют тот же элемент вне системы, т.е., если при исследовании динамики безработицы учесть динамику изменения других социально-экономических показателей, можно обнаружить реальный тренд развития безработицы.

Исходя из основного предположения исследования, была сформулирована задача построения модели, учитывающей влияние различных социально-экономических показателей. Так как взаимное влияние этих показателей в реальной жизни может обнаруживаться не сразу, а через некоторый промежуток времени, в модель был включен временной сдвиг (лаг).

В качестве математического обеспечения решения задачи использовались регрессионный анализ и нейросетевое моделирование.

Метод регрессий является наиболее распространенным подходом к изучению взаимного влияния динамических рядов. Он давно применяется в социологии и достаточно хорошо зарекомендовал себя [6]. Регрессия («regression») означает возвращение в исходное состояние, т.е. отбрасывание случайных ошибок в исходных данных. Наибольшую эффективность регрессионные методы имеют при описании процессов, действующих в достаточно стационарной системе, где отсутствует структурная динамика.

В настоящее время одним из современных направлений в области информатики и вычислительной техники являются нейрокомпьютерные технологии. Их основное достоинство заключается в параллельности вычислений и отсутствии требования стационарности системы, в рамках которой развиваются исследуемые процессы. Известны примеры эффективного использования искусственных нейронных сетей для прогнозирования результатов выборов, значений экономических показателей и анализа данных социологических опросов [7, с. 112–114; 8; 9, с. 36–41].

Как в случае регрессионного анализа, так и в случае нейросетевого моделирования существуют критерии, позволяющие построить для конкретных исходных данных лучшую регрессионную и лучшую нейросетевую модели.

В качестве критерия сравнения регрессионной и нейросетевой моделей использовались их прогнозные возможности. Данные прогноза, полученные посредством моделей, сравнивались с данными Госкомстата.

Исходные данные для построения моделей

Для прогноза динамики безработицы в России, рассчитанной по методике Международной организации труда (МОТ) в процентах, учитывается изменение следующих **показателей:**

1. Занятость (млн. чел.). Российский рынок труда испытывает сильное влияние демографического кризиса. Вместе с безработицей, рассчитанной в процентах, занятость в абсолютных величинах отражает это влияние.

2. Инвестиции в основной капитал (млрд. руб.). Показатель отражает динамику создания рабочих мест.

3. Средняя начисленная заработная плата (руб. в месяц).

4. Потребление товаров и услуг (млрд. руб.).

5. Доход на душу населения (руб. в месяц). В соответствии с методикой МОТ безработным признается человек, ищущий ра-

боту и готовый к ней приступить. Наличие материальных благ, не являющихся результатом трудовой деятельности человека, например, высокая заработная плата других членов семьи, оказывает определенное влияние на принятие человеком решения о необходимости работать.

6. Курс доллара (руб. за долл.). Этот показатель имеет большое влияние на российскую экономику в целом, а следовательно, и на рынок труда.

Далее по тексту номера показателей соответствуют приведенному списку.

При построении моделей учитывалось влияние этих показателей на безработицу без временного сдвига, а также через 3, 6, 9 и 12 месяцев.

В качестве исходных данных были взяты динамические ряды показателя уровня безработицы, рассчитанного по методике МОТ в процентах и перечисленных (шести) показателях. Динамика рассматривалась с января 1996-го по май 2001 г. (шаг дискретизации – 1 месяц) по материалам ежеквартального журнала «Обзор экономики России. Основные тенденции развития», выпускаемого с 1993 г. в соответствии с программой Европейского Союза ТАСИС.

Методика построения регрессионной модели

Первоначально были оценены попарно корреляции динамических рядов (рассматриваемых показателей). Векторы, соответствующие безработице и занятости, оказались слабо коррелированы как между собой, так и со всеми остальными векторами. Коэффициент корреляции между векторами, относящимися к занятости и курсу доллара, составил –0,28, для всех остальных пар векторов, одним из которых являлись занятость или безработица, это значение не превысило по модулю 0,15. Векторы, соответствующие динамическим рядам потребления товаров и услуг, капиталовложений, заработной платы, курса доллара и дохода на

душу населения, попарно дали сильные корреляционные зависимости, лежащие в пределах от 0,71 до 0,99. Однако при сдвиге временного ряда безработицы на 3, 6, 9 и 12 месяцев назад коррелируемость векторов безработицы и занятости как между собой, так и с другими векторами, стала возрастать, превысив по модулю значение 0,5. Это отражает естественную инерционность социально-экономических систем, когда отклик от какого-либо воздействия появляется не сразу, а спустя какое-то время.

Как известно, модель линейной многофакторной регрессии представляет собой уравнение:

$$Y = a_0 + \sum_{i=1}^{n} a_i X_i, \qquad (1)$$

где Y – уровень безработицы, X_i – i-й фактор, учитываемый при ее исследовании, a_i – оцениваемые параметры.

Как показывает практика, число социальных процессов, развивающихся в соответствии с линейной закономерностью, крайне мало. Поэтому перед построением регрессий применялись преобразования исходных данных.

Для построения регрессионных моделей были использованы следующие **преобразования**, нумерация которых употребляется в тексте при анализе результатов: 1 – возведение в куб, 2 – возведение в квадрат, 3 – отсутствие преобразования, 4 – извлечение квадратного корня, 5 – взятие натурального логарифма, 6 – взятие квадратного корня из обратной величины, 7 – взятие обратной величины, 8 – взятие квадрата обратной величины, 9 – взятие куба обратной величины.

Для каждого варианта возможна оптимизация значений коэффициентов a_i для $i = 0, ..., n$ методом наименьших квадратов. Существует множество критериев качества, определяющих, насколько хорошо конкретная форма регрессии описывает исходные данные [10; 11; 12; 13, с. 96–112]. Значение R^2 [10, с. 63] оценивает близость значений Y, полученных из модели, к реальным данным (измеряет объясненную регрессией долю дисперсии Y;

изменяется от 0 до 1; чем ближе значение к 1, тем большая доля дисперсии *Y* объясняется построенной регрессией). Значение *F*-распределения Фишера–Снедекора [11, с. 290] определяет значимость регрессионной модели (если полученное значение больше табличного, то регрессия значима). Значение статистики Дарбина–Ватсона *D-W* [12, с. 220] используется для исследования наборов ошибок (изменяется от 0 до 4, в случае близости этого значения к 2 можно говорить об отсутствии в ошибках какой-либо структуры, т.е. о достаточно хорошей аппроксимации исходных данных).

Если оценить возможное количество регрессионных моделей, которое можно построить, получится внушительное число. У нас имеется шесть показателей, которые мы хотим использовать для анализа динамики безработицы. Над каждым показателем можно сделать 9 преобразований, общее количество возможных регрессионных моделей равно $9^6 = 531441$. Если рассмотреть варианты сдвига динамического ряда безработицы на 3, 6, 9 и 12 месяцев назад, полученное значение нужно умножить на 5 (с учетом варианта без сдвига). И из всех вариантов выбрать наилучший по критериям R^2, *F* и *D-W*. Очевидно, что в такой постановке задача слишком трудоемка.

Для сокращения объема вычислений предложена следующая методика поиска наилучшего вида регрессионной зависимости. Она использует все три критериям качества, но исключает перебор всех возможных вариантов. Главным критерием является критерий Дарбина–Ватсона *D-W*, так как в случае, если он равен 2, простыми средствами улучшить регрессионную модель нельзя. Если значение *D-W* далеко от 2, по критерию Фишера–Снедекора *F* выбираются значимые регрессии, среди которых наилучшими считаются модели, для которых критерий R^2 имеет наибольшее значение.

Исключение необходимости рассмотрения всех возможных регрессионных моделей осуществлено поэтапным вводом в модель показателей. На каждом этапе добавлялся один новый показатель.

На первом этапе были построены регрессионные зависимости безработицы от каждого показателя в отдельности со всеми

возможными преобразованиями. Значения статистики *D-W* для всех регрессий оказались в интервале от 0,04 до 0,6. Значимыми по критерию Фишера–Снедекора для уровня значимости $\alpha = 0{,}01$ оказались лишь регрессии от показателя 1 при преобразованиях 1–7, от показателя 6 при преобразованиях 6–9 и от показателя 3 при преобразовании 1, но значения R^2 для показателя 1 составили 0,74–0,75, а для показателей 6,3 – не превысил 0,22.

На втором этапе рассматривались регрессионные зависимости от двух показателей, одним из которых был показатель 1 (занятость), имеющий в однофакторных моделях наилучшие значения критериев качества.

Далее использовалась аналогичная процедура, т.е. дополнительные показатели вводились только для тех видов регрессии, для которых значения критериев качества *D-W*, *F* и R^2 были наилучшими.

Для принятия решения об изменении преобразования отдельных показателей или удалении их из модели дополнительно использовался *T*-критерий [10, с. 64] – значимость коэффициента при соответствующей переменной (смысл аналогичен *F*-критерию значимости регрессионной модели, но относится не к всему уравнению, а к отдельно взятой переменной из уравнения). В итоге количество рассматриваемых вариантов сократилось до 93. Среди них были выделены 10 наилучших по каждому из трех критериев в отдельности. Дальнейшему исследованию подвергались регрессионные зависимости, попавшие во все три списка, а также наилучшие по отдельным критериям (без учета однофакторных моделей).

По той же методике были построены регрессионные модели по исходным данным со сдвигом динамического ряда уровня безработицы на 3, 6, 9 и 12 месяцев назад.

Наилучшие регрессионные зависимости, как и все остальные, страдали одним существенным недостатком. Значение статистики Дарбина–Ватсона *D-W* ни разу не приблизилось к 2, наилучшее значение было получено на модели, включающей показатели 1 (занятость) и 5 (доход на душу населения) без преобразо-

ваний при сдвиге динамического ряда безработицы на 12 месяцев назад ($D\text{-}W = 0{,}88$). Это объясняется наличием положительной автокорреляции в исследуемых динамических рядах, являющейся следствием малого шага дискретизации (1 месяц). Значение критерия $D\text{-}W < 2$ (наличие положительной автокорреляции [12, с. 229]) объясняется сезонностью. Этот эффект ликвидируется введением фиктивных переменных, которые принимают значение 1 для конкретного месяца года (первая в январе каждого года, вторая – в феврале каждого года и так далее) и значение 0 для всех остальных месяцев. Вводится 11 фиктивных переменных, для декабря каждого года все фиктивные переменные принимают нулевые значения. В противном случае сумма значений фиктивных переменных для любого месяца будет равна 1 и они станут мультиколлинеарны.

Отобранные регрессии были дополнены фиктивными переменными без преобразования.

Желаемый эффект не был достигнут. Для некоторых моделей значение $D\text{-}W$ было незначительно улучшено (но не превысило 1), для других наблюдалось ухудшение.

Следующий этап преобразований преследовал цель ликвидировать возможное влияние как на X_i, так и на Y не учтенных в модели переменных (в том числе теневой экономики). Для этого из списка динамических рядов был выведен доход на душу населения (показатель 5) как редко используемый в процессе построения качественных регрессионных моделей. Для каждого месяца все данные динамических рядов были преобразованы делением на соответствующие данные динамического ряда «доход на душу населения».

Для того чтобы значения преобразованных динамических рядов были одного порядка, они были нормированы. Для показателя 1 использовался коэффициент нормировки 100, для показателя 2 – 1000, для показателя 3 – 10, для показателя 4 – 100, для показателя 6 – 1000, для Y – 1000.

За основу построения регрессионных моделей для преобразованных динамических рядов были взяты наилучшие модели исходных динамических рядов со сдвигом безработицы на 12 месяцев, так как именно они дали наибольшие значения статистики *D-W*.

Для преобразованных динамических рядов без фиктивных переменных наилучшей оказалась модель, включающая показатель 1 (занятость) с преобразованием 7 (взятие обратной величины) и показатель 3 (средняя начисленная заработная плата) без преобразования (*D-W* = 1,6). При добавлении в эту модель фиктивных переменных было получено значение *D-W* = 2,21. При дальнейшем добавлении показателей 4 (потребление товаров и услуг) и 2 (инвестиции в основной капитал) с преобразованиями соответственно 2 (возведение в квадрат) и 1 (возведение в куб) статистика *D-W* приняла свое наилучшее значение 2,07. При этом $R^2 = 0,97$, $F(15,37) = 73,63$, что говорит о том, что ошибки не имеют структуры, регрессионное уравнение объясняет 97% дисперсии *Y* и сама регрессия является значимой (для уровня значимости $\alpha = 0,01$).

Таким образом, качественная регрессионная модель, описывающая динамику безработицы в России, включает преобразование исходных динамических рядов делением данных на соответствующие данные динамического ряда «доход на душу населения», содержит показатель 1 (занятость) с преобразованием 7 (взятие обратной величины), показатель 2 (инвестиции в основной капитал) с преобразованием 2 (возведение в квадрат), показатель 3 (средняя начисленная заработная плата) без преобразования, показатель 4 (потребление товаров и услуг) с преобразованием 1 (возведение в куб) и фиктивные переменные без преобразований, динамический ряд уровня безработицы сдвинут на 12 месяцев назад. Для получения прогнозных значений уровня безработицы из регрессионной модели данные были подвергнуты обратной нормировке.

Методика построения нейросетевой модели

Под нейронными сетями подразумеваются вычислительные структуры, которые упрощенно моделируют биологические процессы человеческого мозга. Они представляют собой распределенные и параллельные системы, способные к адаптивному обучению. Элементарным преобразователем в данных сетях является искусственный нейрон, названный так по аналогии с биологическим прототипом. К настоящему времени предложено большое количество моделей нейроподобных элементов и нейронных сетей [8].

Для анализа уровня безработицы был использован нейросетевой пакет NeuroPro [8, с. 169–188], свободно распространяемый в Интернете. Основными достоинствами пакета являются русскоязычность, простота в изучении, возможность упрощения сети и выявления наиболее (или наименее) значимых входов.

NeuroPro позволяет создавать многослойные нейронные сети. Каждый слой состоит из нескольких нейронов. Каждый нейрон слоя принимает на себя сигналы от всех нейронов предыдущего слоя сети и передает сигнал на каждый нейрон последующего слоя. Первый слой является входным и соответствует независимым показателям X_i (занятости, инвестициям в основной капитал, средней начисленной заработной плате, потреблению товаров и услуг, доходу на душу населения, курсу доллара). Последний слой является выходным и соответствует показателю Y (уровень безработицы). В пакете используется следующая формула нейрона:

$$Y = f(s) = \frac{s}{|s| + c}, \qquad (2)$$

где $s = \sum_{i=1}^{n} w_i x_i + b$, x_i – показатель входного вектора (входной сигнал) или сигнал от нейрона предыдущего слоя, w_i – коэффициент (вес), определяемый в процессе настройки сети (вычисляется на

основе значений входных и выходных данных), b – значение смещения, c – характеристика, определяющая крутизну наклона функции $f(s)$, задается для каждого слоя отдельно.

В используемой версии NeuroPro реализован алгоритм обратного распространения ошибки. После настройки весов w_i (процедура в определенной степени напоминает вычисление коэффициентов a_i в регрессионной модели) теоретические значения Y сравниваются с реальными значениями выходного слоя (уровень безработицы) и если ошибка превышает заданную, w_i пересчитываются в соответствии с итеративным градиентным алгоритмом. В пакете реализованы четыре модификации алгоритма: градиентный спуск, модифицированный партан, сопряженные градиенты и BFGS – разновидность оптимизационного алгоритма Ньютона.

Как известно, увеличивая количество промежуточных слоев и нейронов в них, можно обеспечить практически любую точность вычисления Y. Однако это не будет являться гарантией качества построенной сети. Для контроля способности сети описывать реальность, часть данных резервируется в «экзаменационную» (контрольную) выборку. Данные, по которым осуществляется вычисление весов w_i, соответственно, образуют обучающую выборку. Обучение сети заканчивается в случае достижения на обучающей выборке заданной точности вычисления Y, либо в случае невозможности дальнейшего сокращения ошибки. Качество сети проверяется по ошибке, которую дают теоретические значения Y на контрольной выборке. Наилучшей считается сеть, которая дает минимальную ошибку на контрольной выборке.

При построении нейросетевой модели уровня безработицы в качестве экзаменационной выборки резервировалось каждое четвертое значение динамических рядов.

По умолчанию NeuroPro предлагает сеть, состоящую из трех слоев по 10 нейронов в каждом. Кроме сетей этой структуры, при анализе уровня безработицы рассматривались двухслойные сети. Для них имеются формулы, оценивающие количество нейронов в слоях

[8, с. 22]. Количество нейронов в каждом слое зависит от числа входов, выходов, элементов обучающей выборки и определяется двумя значениями – минимально и максимально необходимым количеством нейронов в слое. При построении двухслойных сетей использовалось максимально необходимое количество нейронов в слое.

Для анализа динамики безработицы были построены нейронные сети, созданные последовательно для наборов динамических рядов без сдвига и со сдвигом ряда безработицы на 3, 6, 9 и 12 месяцев. Выходом каждой сети являлся динамический ряд безработицы, на входе – шесть показателей, от которых, по нашему мнению, зависит уровень безработицы. Для каждого варианта сдвига динамического ряда безработицы созданы 8 сетей – 4 двухслойной и 4 трехслойной структуры (по числу возможных алгоритмов обучения). Каждый алгоритм протестирован на зарезервированной выборке. Ошибка, задаваемая для обучения сети, составила 0,1 с последующим уменьшением до 0,01. В большинстве случаев уменьшение ошибки приводило к ухудшению аппроксимационных возможностей сети. Качество сети определялось единственным показателем – максимальной ошибкой, даваемой сетью на резервной (контрольной) выборке.

Для каждого набора динамических рядов (без временного сдвига и со сдвигами ряда уровня безработицы) в структуры сетей, имеющих наименьшие ошибки для резервной выборки, были добавлены фиктивные переменные на входе аналогично фиктивным переменным в регрессионных моделях.

Наилучшими аппроксимационными возможностями обладала сеть, построенная при сдвиге динамического ряда уровня безработицы на 9 месяцев назад, с включением фиктивных переменных. Нейронная сеть состояла из 3 слоев по 10 нейронов в каждом, при задаваемой точности обучения 0,1. Максимальная ошибка, полученная на резервной (экзаменационной) выборке, составила 0,6. Для каждой нейронной сети определялась значимость входных сигналов (сила влияния на выходной сигнал). Она

изменяется от 0 до 1 и тем выше, чем ближе к 1. Для нейронной сети, давшей наилучший результат на экзаменационной выборке, значимость показателя 1 (занятость) составила 1, показателя 2 (инвестиции в основной капитал) – 0,08, показателя 3 (средняя начисленная заработная плата) – 0,11, показателя 4 (потребление товаров и услуг) – 0,34, показателя 5 (доход на душу населения) – 0,36, показателя 6 (курс доллара) – 0,43, значимость фиктивных переменных – в пределах от 0,16 до 0,48.

Прогнозные возможности построенных моделей

Прогнозные значения уровня безработицы на пять месяцев вперед, полученные из наилучших регрессионной и нейросетевой моделей, были сравнены с данными Госкомстата, рассчитанными по результатам выборочного обследования населения.

Максимальная ошибка, полученная на наилучшей регрессионной модели, составила в абсолютном выражении 4,98 (более 60% реального значения уровня безработицы). Нужно отметить, что регрессионная модель, обладающая несколько худшим значением качества (значение $D\text{-}W = 1{,}88$ по сравнению с $D\text{-}W = 2{,}07$ для лучшей модели), полученная включением в наилучшую модель показателя 6 (курс доллара) с преобразованием 2 (возведение в квадрат), дала меньшую прогнозную ошибку (максимальная абсолютная ошибка 3,14 или 38,29%).

Как уже отмечалось выше, качественная регрессионная модель хорошо описывает процессы, происходящие в системе, не имеющей ярко выраженной динамики структурного развития. Этим можно объяснить тот факт, что наилучшая с точки зрения формальных критериев качества регрессионная модель приводит к лучшему прогнозу изменения численности безработных даже по сравнению с другими регрессионными моделями, что затрудняет возможность использования регрессионного анализа для прогноза социально-экономических процессов, происходящих в России в настоящее время.

В рамках «лучшей» нейросетевой модели была получена наибольшая абсолютная ошибка 1,33, что составляет 16% реального значения уровня безработицы.

Нейросетевое моделирование показало существенные преимущества перед регрессионным анализом, основными из которых являются меньший объем необходимых вычислений для построения модели и большая точность результатов.

ЛИТЕРАТУРА

1. *Гильдингерш М.Г.* Безработица в России: сущность, формы, социальные последствия в условиях перехода к рынку: Автореф. дис. ... д-ра эконом. наук. СПб, 1995.

2. *Пошевнев Г.С.* Регулирование занятости и безработицы как функция социального управления: Автореф. дис. ... д-ра социол. наук. М., 1997.

3. *Прокопов Ф.Т.* Безработица и эффективность государственной политики труда в переходной экономике России: Автореф. дис. ... д-ра эконом. наук. М., 1999.

4. *Волконский В.А., Корягина Т.И.* Официальная и теневая экономика в реальности и статистике // Экономика и математические методы. 2000. Т. 36. № 4.

5. *Николаенко С., Лиссовик Я., Мак-Фаркар Р.* Теневая экономика в российских регионах // Обзор экономики России. 1997. IV.

6. *Толстова Ю.Н.* Анализ социологических данных: Методология, дескриптивная статистика, изучение связей между номинальными признаками. М.: Научный мир, 2000.

7. *Круглов В.В., Дли М.И.* Применение аппарата нейронных сетей для анализа социологических опросов // Социологические исследования. 2001. № 9.

8. *Круглов В.В., Борисов В.В.* Искусственные нейронные сети: Теория и практика. М.: Горячая линия – Телеком, 2001.

9. *Молоканов В.Д., Долганов А.П., Секерин А.Б.* Использование технологии нейронных сетей для прогнозирования налоговых поступлений на основе унифицированной системы показателей Госстатистики // Вопросы статистики. 2000. № 7.

10. *Глинский В.В., Игонин В.Г.* Статистический анализ. М.: Филинъ, 1998.

11. *Гмурман В.Е.* Теория вероятностей и математическая статистика. 7-е изд., стер. М.: Высшая школа, 2000.

12. *Доучерти К.* Введение в эконометрику. М.: ИНФРА–М, 1999.

13. *Крыштановский А.О.* Ограничения метода регрессионного анализа // Социология: методология, методы, математические модели. 2000. № 12.