Анна Вайнберг Аллен

Графы для анализа структурных соотношений между переменными и их приложение к изучению российских регионов (часть 2)

Вторая часть статьи продолжает исследование структуры набора случайных переменных. Она состоит из двух частей: 1) описание предлагаемой автором модификации метода выбора ковариаций Демпстера, основанной на его комбинации с алгоритмом построения деревьев зависимостей, результаты моделирования, а также технология представления данной графовой модели на плоскости и различные методы интерпретации результатов; 2) применение разработанного метода к практическому исследованию и сравнению российских регионов.

В разделе 4 подробно описаны модификация алгоритма Демпстера и связанные с ней методы моделирования и интерпретации. Все программы, представленные в данной работе, и примеры их применения можно найти на сайте stat.solev.ru/weinberg.

Раздел 5 посвящен применению модифицированного алгоритма Демпстера к анализу российских регионов в 1994–1999 годах. Мы наблюдаем своего рода «поле переменных» — плавный переход от переменных, характеризующих экономическую, предпринимательскую деятельность и качество населения, к общим макроэкономическим индикаторам (ВРП и др.), а затем через инфраструктурные и географические индикаторы — к социальным индикаторам. Такое «поле» также содержит изолированные переменные.

Основное внимание уделено анализу графов за 1994, 1997 и 1999 годы, в частности анализу структуры переменных и ее изменению во времени.

Понимание структуры переменных позволяет выяснить, какие внешние воздействия (например, развитие ипотеки или рост промышленного производства) могли бы привести к наиболее ощутимым практическим социально-экономическим результатам. С точки зрения экономической теории особый интерес вызывает сравнение структур переменных для различных наборов данных.

Автор выражает свою глубокую благодарность проф. С. А. Айвазяну за постоянную поддержку в процессе написания этой статьи, а также проф. Ж. Антилю, проф. Ю. Н. Благовещенскому, своему отцу с.н.с. Л. И. Вайнбергу и с.н.с. Т. С. Рыбниковой за полезное обсуждение и ценные советы. Автор также хотел бы отметить, что эта работа была начата в 1999 году совместно с ныне покойным проф. Л. Д. Мешалкиным.

4. Новый алгоритм выбора

Опишем модификацию алгоритма Демпстера [Dempster (1972)], детально представленного в первой части статьи. Главная идея состоит в следующем:

1) объединение дерева зависимостей и модели выбора ковариаций, введенных в первой части статьи;

2) разработка дополнительных инструментов для интерпретации результатов.

Как ранее обсуждалось в подразделе 1.3, I[1], мы часто предполагаем, что *гипотеза древовидных зависимостей* удовлетворяется для определенного подмножества данных. Таким образом, мы идентифицируем начальную субмодель, выбирая первые ребра дерева зависимостей. Эта модель используется как начальная точка в алгоритме Демпстера. Как правило, она уже близка к решению, и такой выбор уменьшает число вычислений.

Другое преимущество выбора дерева зависимостей в качестве начальной точки заключается в использовании простой и понятной структуры. Действительно, идея дерева зависимостей берет свое начало в цепях Маркова и их свойстве условной независимости будущего от прошлого.

Алгоритм выбора ковариаций Демпстера также основан на условной независимости между переменными. Однако итерационный алгоритм в целом труднее для понимания и построенные графы имеют более сложную структуру.

4.1. Описание алгоритма

Алгоритм в целом можно описать следующим образом:

Шаг инициализации. Вычислить дерево зависимостей (алгоритм Крускала). Вычислить логарифмическую функцию правдоподобия диагональной корреляционной матрицы.

Шаг I. Добавить ребро дерева зависимостей. По алгоритму Дейкстры оценить корреляционную матрицу, используя свойство цепи деревьев зависимостей в случае нормального распределения (см. Приложение 1).

Шаг II. Если это ребро вносит достаточный вклад[2], то оно добавляется, и происходит возврат к шагу I.

Шаг III. Применить алгоритм Демпстера выбора ковариаций (см. раздел 3, I).

Конец алгоритма.

В результате получаем «улучшенное» дерево зависимостей, т. е. усеченную древовидную структуру зависимостей с дополнительными ребрами, или, если посмотреть с другой стороны, модификацию алгоритма Демпстера.

В данном случае результат выполнения алгоритма Крускала служит отправной точкой для реализации алгоритма Демпстера. Теоретическая сложность алгоритма остается неизменной, однако использование новой начальной точки заметно уменьшает число вычислений.

4.2. Псевдокод

Технически новый алгоритм выбора состоит из последовательного применения трех алгоритмов:

[1] Здесь и далее римской цифрой I обозначена первая часть статьи, к которой принадлежит соответствующий раздел, рисунок или алгоритм. Например, 1.3, I — это подраздел 1.3 первой части статьи.

[2] Вклад считается достаточным, если значима разность между новым и старым значением критерия, построенного на основе логарифмической функции правдоподобия (см. правило остановки в разделе 3, I).

1) **алгоритма Крускала** поиска максимального связывающего дерева (MST) на шаге инициализации;

2) **алгоритма Дейкстры** [Dijkstra (1959)] поиска кратчайшей траектории на шаге I (алгоритм 3, см. Приложение 1);

3) **алгоритма выбора ковариаций Демпстера** на шаге III (алгоритмы 1 и 2, I).

4.3. Численный пример

Вернемся к примеру Демпстера. Для этого примера с данными, имеющими древовидную структуру зависимостей, все три алгоритма, изложенные в данной статье (дерево зависимостей, алгоритм Демпстера и модифицированный алгоритм Демпстера), приводят к одному и тому же результату, изображенному в виде графа на рис. 1, I.

Шаг инициализации состоит в построении дерева зависимостей, и он представлен в численном примере подраздела 3.3, I.

Шаги I–II

Итерация 1: Добавляем первое ребро дерева зависимостей: (4, 5). Увеличение значения функции правдоподобия незначимо.

Итерации 2–4: Добавляем следующие три ребра согласно алгоритму Крускала.

Итерация 5: Добавляем пятое ребро дерева зависимостей: (5, 6). Корреляционная матрица, оцененная по алгоритму Дейкстры, имеет вид

$$\hat{\Sigma} = \begin{pmatrix} 1{,}000 & 0{,}3966 & 0{,}3688 & 0{,}2163 & -0{,}4632 & 0{,}1693 \\ & 1{,}0000 & 0{,}1463 & 0{,}0858 & -0{,}1837 & 0{,}0672 \\ & & 1{,}0000 & 0{,}0798 & -0{,}1708 & 0{,}0625 \\ & & & 1{,}0000 & -0{,}4671 & 0{,}1708 \\ & & & & 1{,}0000 & -0{,}3656 \\ & & & & & 1{,}0000 \end{pmatrix}.$$

Разность значений логарифмических функций правдоподобия равна 28,32. Значение критической статистики, основанной на логарифмической функции правдоподобия, значимо.

Таким образом, **шаг III** не выполняется.

4.4. Моделирование

В разделе 5 модифицированный алгоритм Демпстера будет применен к реальным сложным совокупностям данных. Поэтому вначале тестируем поведение обоих алгоритмов — алгоритма выбора ковариаций Демпстера и модифицированного алгоритма Демпстера — с помощью смоделированных данных.

4.4.1. Моделирование данных

Моделируем нормально распределенные совокупности данных X_{ij}: $i = 1,\ldots,n$; $j = 1,\ldots,p$, где n — число наблюдений, p — число переменных. Фиксируем $n = 100$, поскольку на практике как в межстрановых, так и в межрегиональных исследованиях никогда не бывает большего числа наблюдений, и p изменяется от 3 до 25. В частности, моделируем данные:

- случайные;
- с древовидной структурой зависимостей;
- с частичной древовидной структурой зависимостей;
- с блочно-диагональной структурой корреляционной матрицы.

4.4.2. Моделирование корреляционных матриц с известной структурой

Процедура моделирования — одна и та же для всех четырех случаев. Моделируем корреляционную матрицу C с определенной структурой. Для этого прежде всего строим матрицу случайных величин $V = \{V_{ij}\}$, $i = 1,\dots,n$; $j = 1,\dots,p$, или, другими словами, p векторов $V_{\cdot j}$. Далее находим $\hat{C} = 1/n\,(V - \bar{V})'(V - \bar{V})$.

Затем, применяя разложение Холецкого, получаем матрицу H, такую, что $\hat{C} = H'H$, и строим нормально распределенные данные X, применяя формулу $X = YH$, где $Y_{n\times p} \sim N[0, I]$.

Опишем способы моделирования V_{ij} для четырех выбранных структур данных.

I. Случайные данные. Моделируем случайную матрицу $V_{ij} \sim N(0, 1)$, $i = 1,\dots,n$; $j = 1,\dots,p$, где V_{ij} взаимонезависимы. Получаемая выборочная ковариационная матрица $\hat{C}$ должна быть близка к I.

II. Древовидная структура зависимостей. Вначале моделируем $V_{i1} \sim N(0, 1)$ с независимыми координатами, далее строим вектор $V_{\cdot j+1}$ на основе вектора $V_{\cdot j}$ согласно следующей процедуре:

1) моделируем $\alpha_j \sim R(0{,}1; 0{,}3)$[3] и берем число $l_j = \alpha_j n$, округленное до следующего целого;
2) моделируем константу c_j, где $c_j \sim R(-1{,}5; 1{,}5)$;
3) случайным образом выбираем l_j элементов вектора $V_{\cdot j}$ и умножаем каждый из них на константу c_j, остальные его элементы остаются без изменения. Полученный вектор называем $V_{\cdot j+1}$.

III. Частичная древовидная структура зависимостей. Строим k блоков с m (полагаем $m = 5$) векторами в каждом (кроме, возможно, последнего блока), где k равно отношению числа переменных p к m, округленному до следующего целого числа.

Повторяем процедуру II отдельно для каждого блока, но при этом α_j выбирается из $R(0{,}1; 0{,}5)$, а c_j — из $R\big((-2; -4)\bigcup(2; 4)\big)$.

IV. Блочно-диагональная структура. Строим блоки, как в случае III. Далее моделируем первый вектор в каждом блоке, как в случае II, и строим вектор $V_{\cdot j+1} = V_{\cdot j} + U_{\cdot j}$, где координаты $U_{\cdot j}$ — независимые случайные величины из $N(0; 0{,}2)$.

4.4.3. Результаты моделирования

Полученные результаты очень близки для всех четырех смоделированных структур данных. Для экономии места приведем результаты только для данных с частичной древовидной структурой зависимостей, поскольку именно эта структура всегда встречалась нам в практических исследованиях.

[3] Через $R\,(a; b)$ обозначаем равномерно распределенную величину на отрезке $(a; b)$.

Алгоритм Демпстера и модифицированный алгоритм Демпстера практически всегда приводят к одному и тому же результату, но их эффективность отличается.

На рис. 3 для частичной древовидной структуры зависимостей показано, что абсолютная разность времен выполнения алгоритмов Демпстера увеличивается при увеличении числа переменных. Это верно и для других рассматриваемых структур данных.

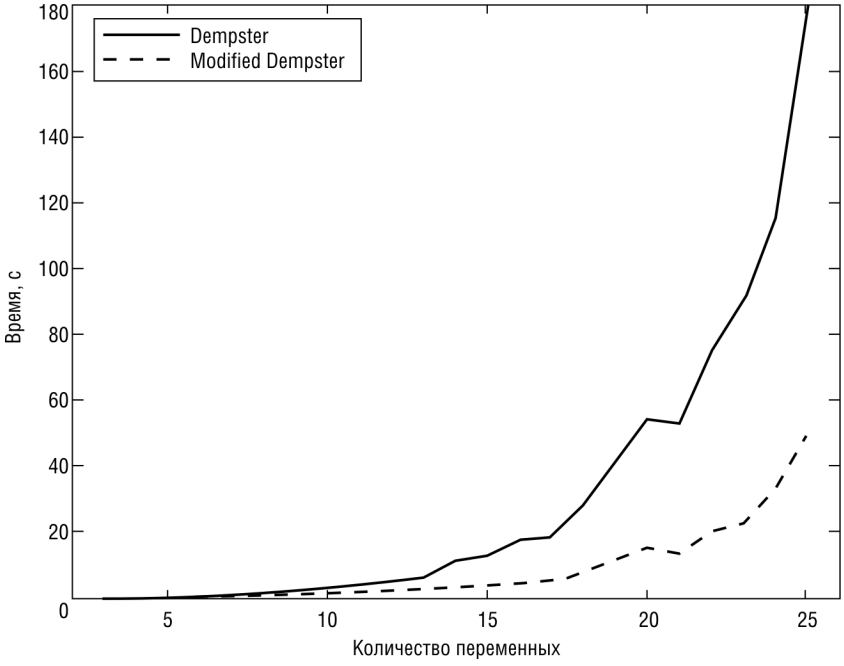

Рис. 3. Время выполнения алгоритма при различном числе переменных для данных с частичной древовидной структурой зависимостей

Однако нас больше интересует скорость этого увеличения. Для частичной древовидной структуры зависимостей и случайных данных на рис. 4 представлены отношения времени выполнения двух алгоритмов. Видно, что для данных с частичной древовидной структурой зависимостей (см. рис. 4, *а*) при увеличении числа переменных отношение времени выполнения алгоритмов стабилизируется около цифры 4, а для случайных данных (см. рис. 4, *б*) — около цифры 2, так же как и для двух других рассматриваемых структур данных.

Напомним, что модифицированный алгоритм Демпстера специально разработан при предположении частичной древовидной структуры зависимостей данных. Таким образом, кажется логичным, что он более эффективен именно в присутствии этой структуры данных.

Преимущество во времени выполнения не настолько ощутимо, но оно, тем не менее, заметно в практической работе, особенно принимая во внимание, что совокупности данных в пространственном (межрегиональном) анализе страны, как правило, имеют частичную древовидную структуру зависимостей.

На рис. 5 показано интересное наблюдение для частичной древовидной структуры зависимостей: число связей, сохраненных алгоритмом, в основном пропорционально числу переменных. Это верно и для трех других моделируемых структур данных.

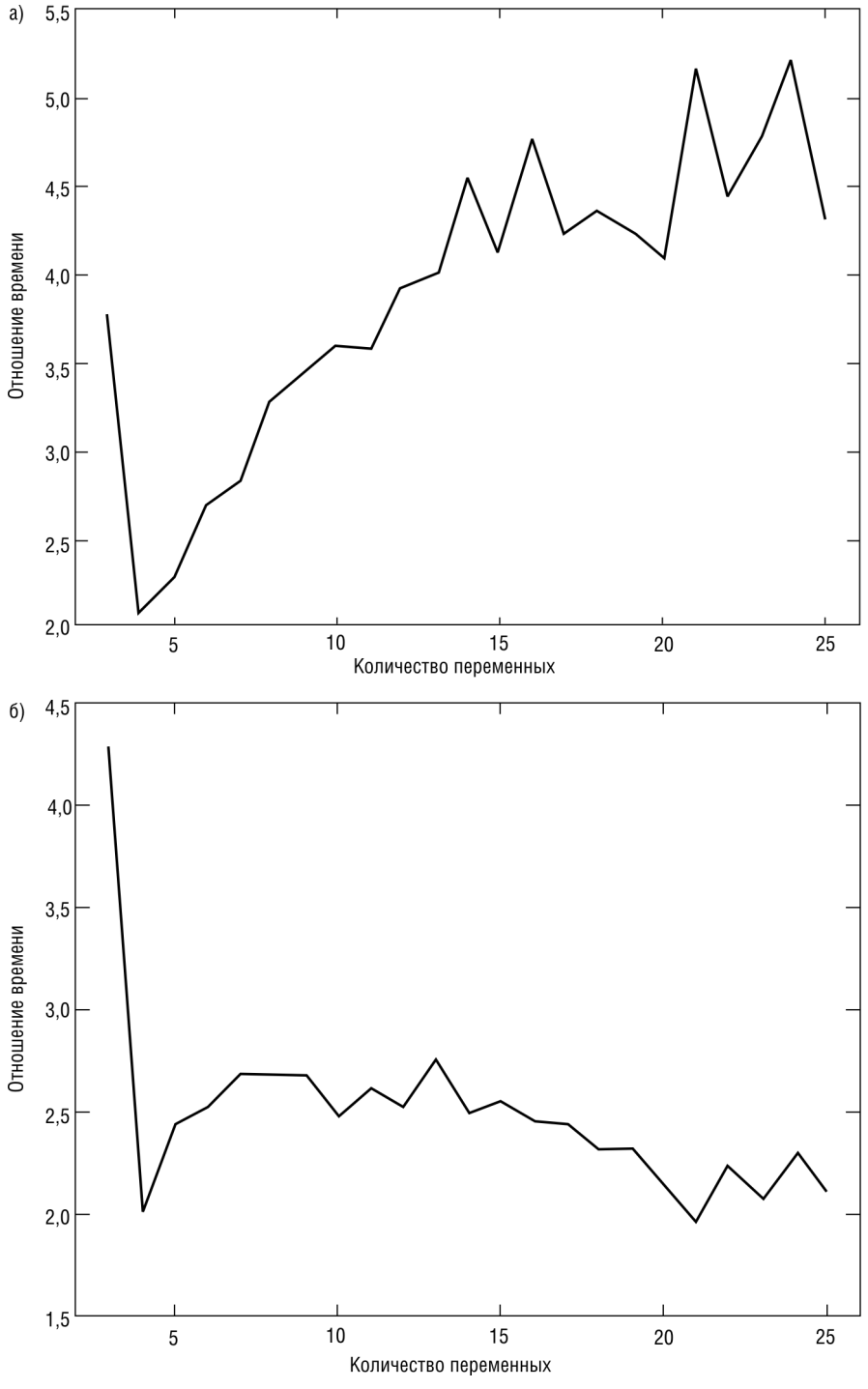

Рис. 4. Отношение времени выполнения алгоритма Демпстера и модифицированного алгоритма Демпстера для данных с частичной древовидной структурой зависимостей (*а*) и для случайных данных (*б*)

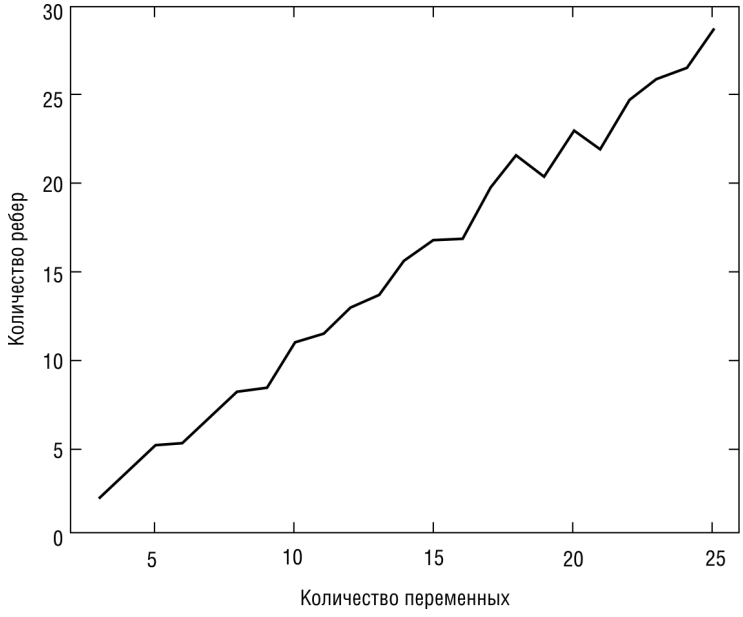

Рис. 5. Соотношение между числом ребер и числом переменных для данных с частичной древовидной структурой зависимостей

4.5. Процедура построения графа

Для построения графовых моделей используем систему визуализации графов uDraw(Graph)[4], разработанную в Университете Бремена. Система распространяется бесплатно для использования в исследовательских целях и доступна непосредственно на сайте uDraw (Graph)[5].

Данные можно вводить вручную либо они должны быть подготовлены в формате API (Application Programmer Interface) — внутреннем языке uDraw(Graph). Нами в системе MatLab написан специальный интерфейс, позволяющий переводить графы, представленные в виде троек [вершина 1, вершина 2, вес ребра] в формат API. Напоминаем, что мы работаем с неориентированными графами. Программа интерфейса находится в свободном доступе на сайте stat.solev.ru/weinberg.

4.6. Интерпретация

Обсудим три аспекта интерпретации результатов: качество объяснения, интерпретация структуры переменных и интерпретация наблюдений.

4.6.1. Качество объяснения

Применяем два индикатора для оценивания качества объяснения:

- *качество представления корреляционной матрицы* (см. подраздел 1.2, I). Этот индикатор отражает, насколько хорошо из оставленных в графе ребер, можно восстановить исключенные ребра исходной выборочной корреляционной матрицы;

[4] До 2005 года программа была известна как daVinci или da Vinci Presenter.

[5] www.informatik.uni-bremen.de/uDraw(Graph)/en/index.html

• *доля* (*в процентах*) *логарифмической функции правдоподобия* (далее для краткости будем говорить: логарифм правдоподобия), *объясняемая графом*. Вычисляем отношение логарифма правдоподобия для матрицы, аппроксимированной графовой моделью, к логарифму правдоподобия корреляционной матрицы.

4.6.2. Интерпретация структуры переменных

Опишем основные методы и индикаторы, применяемые для интерпретации структуры переменных.

1. Основная идея заключается в **непосредственной интерпретации графа**, чтобы распознать структуру корреляционной матрицы, обнаружить переменные, которые агрегируются в другие переменные, и отследить взаимосвязи между переменными и группами переменных.

В частности, можно провести сравнительное исследование различных графов для одного и того же множества переменных (но для различных групп регионов, стран или для различных временных диапазонов), изучая поведение и устойчивость во времени и в пространстве структуры переменных.

2. Вводим также **иерархию переменных**. Их можно упорядочить, во-первых, по суммам логарифмов правдоподобия исходящих ребер, во-вторых, по числу связей (числу смежных вершин) для каждой переменной и, в-третьих, согласно некоторой функции от этих двух переменных.

3. Большинство наших выводов получаем, **интерпретируя множество самых «информативных» переменных**, другими словами, тех, которые имеют высокие значения логарифма правдоподобия и(или) большое количество связей. Определение 7 из первой части статьи вводит понятие степени переменной: степень переменной равняется числу ее связей. Переменные третьей степени и выше мы называем *скелетными*. Полагаем, что такие переменные в некотором смысле являются агрегацией других, связанных с ними переменных.

В частности, для каждой переменной i вычисляются следующие итоговые параметры:

а) число смежных вершин n_i;

б) функция от значений логарифмов правдоподобия ребер, выходящих из данной вершины. Вычисляем также sum_i — сумму этих значений;

в) оптимизационная функция — комбинация первого и второго параметров. Обозначим через $\mathrm{sum}_{\max}$ максимальную сумму значений логарифмов правдоподобия смежных ребер, а через $n_{\max}$ — максимальное число связей. Таким образом, данная комбинация для переменной i определяется как $\mathrm{sum}_i + \mathrm{sum}_{\max} \cdot n_i / n_{\max}$.

4. **Значение R^2 для регрессии** вычисляется только для переменных, имеющих смежные ребра[6]. Эта переменная выступает в качестве независимой, а ее смежные вершины — в качестве объясняющих переменных. Большое значение R^2 означает, что смежные переменные допускают хорошую аппроксимацию самой переменной.

[6] В противном случае этот параметр устанавливается равным нулю.

5. **Качество представления графа и отдельных переменных**. Хорошее качество представления означает, что на основе ребер, входящих в граф, можно достоверно восстановить все корреляции данной переменной.

Оптимизационная функция (комбинация) между числом смежных ребер и суммой значений логарифмов правдоподобия особенно полезна, когда множество переменных содержит подмножество высококоррелированных переменных. Ребра, соединяющие переменные этого подмножества, добавляются первыми. Таким образом, соответствующие переменные имеют особенно высокие суммы значений логарифмов правдоподобия. Относительная важность других переменных при этом занижается. В частности, в нашем случае такой эффект наблюдается для экономических переменных. Последние тесно связаны между собой, и поэтому мы начинаем с добавления переменных, которые имеют высокие значения логарифмов правдоподобия.

Таким образом, используя только значения логарифмов функции правдоподобия, вводим в наш анализ «экономическое смещение», а дополняя логарифм правдоподобия степенью переменной, в значительной мере «корректируем» этот нежелательный эффект.

4.6.3. Интерпретация наблюдений

Мы можем объединить исследование переменных с анализом (типологией) наблюдений. В данной статье, в связи с ограничением по объему, мы не имеем возможности представить конкретные примеры интерпретации, но опишем сами методы.

1. Для каждого наблюдения строим «звезду Велша» (Welsh) [Fienberg (1979)], что приводит к полезному графическому представлению данных. Для каждого наблюдения значения переменных откладываются на равноотстоящих друг от друга радиусах, исходящих из центра круга, формируя своего рода «звезду». Если переменные измерены в разных единицах, то они стандартизируются перед построением графика. Можно построить звезды, используя все переменные множества или только скелетные, подмножество наиболее информативных переменных или подмножество взаимосвязанных переменных.

2. Вместо звезд Велша можно использовать «лица Чернова» (Chernoff). Лица Чернова аналогичны лицам людей, и, таким образом, большее значение приобретают те переменные, которые соответствуют более заметным (изолированным) чертам лица. Для эффективного «чтения» лиц Чернова важен порядок, в котором переменные ассоциируются с определенными чертами лица, и знак этих переменных. Как и в предыдущем случае, переменные стандартизируются, и дополнительно мы обращаем внимание, чтобы «лучшим» характеристикам региона соответствовали положительные знаки переменных. Области с «более высоким развитием» имеют «более удовлетворенное выражение лица».

3. К выбранному множеству переменных применяем также кластерный анализ. Используем результаты иерархических методов [Johnson (1967), Borgatti (1994)], чтобы распознать структуру наблюдений и выбрать начальную точку. Далее, для уточнения разбиения, применяем метод динамических облаков с устойчивыми ядрами [Diday (1971), Diday et al. (1982), Ammor, Chah Slaoui (2000)]. Могут использоваться также и другие методы кластерного анализа. Однако метод динамических облаков с устойчивыми ядрами позволяет автоматически определить число кластеров данного множества переменных. Более подробное описание

метода и программу, написанную нами в системе MatLab, можно найти на сайте stat.solev.ru/weinberg.

Наш опыт применения кластерного анализа показывает, что наиболее интересные результаты достигаются при построении кластеров на основе подмножества взаимосвязанных переменных, отражающих не более двух «глобальных» характеристик переменных.

4. Множество скелетных переменных также имеет интересное свойство «сохранения картинки». Для ряда наборов данных мы сравнивали плоскости первых двух главных компонент, построенные: 1) на самом наборе переменных и 2) только на основе переменных со степенями более двух при сохранении того же набора наблюдений. Расположение наблюдений остается почти тем же самым. Визуально создается впечатление, что местоположение регионов относительно друг друга остается очень похожим. В будущем представляет интерес проверить данное наблюдение на статистически значимом количестве наборов данных и ввести численные критерии сравнения.

4.7. Направления продолжения работы

Ниже представлен неполный перечень дополнительных возможностей для применения и интерпретации графовых моделей (часть из них реализована в данной работе).

1. Можно строить графы для отдельных больших групп переменных, например для экономических, социально-демографических, политико-правовых и экономико-правовых переменных. На предварительном, «разведочном» этапе исследования, когда подчас приходится выбирать из сотен переменных, это имеет смысл сделать с помощью деревьев зависимостей, скорость построения которых почти не зависит от количества переменных.

2. Можно повторить исследование с тем же самым множеством переменных в разные периоды времени или для различных подмножеств наблюдений (в этой работе, например, изучаются отдельно регионы Сибири и европейской части России в данный момент времени).

3. Можно расширить набор используемых переменных, преобразуя переменные, изначально не распределенные нормально, в нормально распределенные (более подробно об этом см. Приложение 3.3).

4. Можно (и до́лжно!) проверить с помощью теста Грейнджера направленность каждого ребра графа, т. е. какие факторы являются причиной, а какие — следствием той или иной связи.

5. Иногда можно объединить наблюдения за различные периоды времени. Допустим, имеется два набора данных по российским регионам, например, за 1996 и 2006 год, где каждый набор содержит 77 наблюдений. Тогда в объединенном наборе будет 144 наблюдения с именами, состоящими из названия региона и года, к которому относится указанный набор данных. В нашем случае наблюдения будут с именами «Москва 1996», «Москва 2006», «Татарстан 1996» и «Татарстан 2006». Данная процедура имеет смысл, когда не хватает наблюдений и(или) когда происшедшие изменения кардинально меняют сам объект наблюдения. Например, в нашем случае можно определенно сказать, что Москва в 1996 и 2006 году — это экономически и социально два совершенно разных общества.

6. В будущем можно применить результаты, полученные с помощью графовых моделей, для построения эконометрических моделей. В этом случае строим уравнения регрессии для скелетных переменных и рассматриваем переменные, с которыми они связаны в качестве объясняющих. При этом получаем систему одновременных уравнений.

5. Анализ российских регионов в 1994–1999 годах

Данный раздел посвящен практическому применению графовых моделей (модифицированного алгоритма Демпстера) к исследованию российских регионов.

Традиционно, с помощью метода анализа главных компонент получаем новые переменные как агрегации других переменных. Применение графовых моделей позволяет подойти к решению задачи с другой стороны: мы находим переменные, которые уже являются агрегациями других переменных. В анализе главных компонент ищем переменные (направления), максимизирующие дисперсию всех «облаков» переменных. Применяя алгоритм Демпстера, как описано в разделе 3, I, выбираем ребра, суммирование которых обеспечивает максимизацию информации, содержащейся в других ребрах, и, таким образом, максимизацию логарифма правдоподобия.

В данном исследовании проводим анализ 77 многомерных наблюдений, представлявших регионы — субъекты Российской Федерации. Автономные округа, а также республики Ингушетия и Чечня были исключены из исследования. Набор данных покрывает шестилетний период с 1994 по 1999 год и содержит 29 переменных (см. Приложение 2).

5.1. Выбор переменных

Для проведения нашего исследования необходимо было охватить как можно больше различных аспектов ситуации в регионах, а также их развитие. В любом межрегиональном или межстрановом исследовании Моррис и Адельман [Morris, Adelman (1988)] предлагают использовать пять групп переменных:

E — *экономические индикаторы*;
D — *демографические индикаторы*;
S — *социально-институциональные переменные* и *переменные, характеризующие человеческий капитал*;
P — *политико-институциональные переменные*;
M — *рыночно-институциональные переменные*.

Такая классификация переменных вполне пригодна для наших целей, однако мы ввели еще одну группу:

G — *географические индикаторы*,

которая отражает российскую специфику: протяженность и разнообразие территории страны.

Выбор переменных во всех шести группах в первую очередь определяется нашим намерением использовать интегрированные индикаторы, полностью характеризующие некоторое количество аспектов социально-экономической ситуации. Также выбор переменных обусловлен желанием проверить ряд социально-экономических гипотез: о роли инвестиций, человеческого капитала, природных ресурсов, предпринимательской активности, географического положения регионов и институциональных факторов. Примером интегрированного индикатора, в частности, служит ВРП на душу населения — переменная, отражающая целый ряд аспектов экономической ситуации в регионе. Вместе с тем во многих случаях интегрированные индикаторы оказались недоступны, поэтому мы были вынуждены заменить их другими. Например, в нашем распоряжении не было комплексного индикатора климатических условий в российских регионах. Вместо него была использована средняя темпе-

ратура января, причем для того, чтобы зафиксировать этот индикатор, пришлось использовать одну и ту же температуру для всех лет на протяжении изучаемого периода.

Кроме того, мы не располагали переменными, описывающими институциональное развитие, а также действия правительства и проведение реформ, поэтому при изучении политико-институциональных характеристик регионов были вынуждены использовать результаты выборов.

Более подробную информацию по указанному набору данных, включая проверку его на нормальность, преобразование отдельных переменных, работу с пропущенными переменными, можно найти в Приложении 3. Строим графовые модели за все 6 лет изучаемого периода, а затем исследуем стабильность и изменения переменных.

Для 1994 года представлен полный набор методов интерпретации. Список ребер с соответствующими корреляционными коэффициентами и значениями логарифмов правдоподобия содержится в табл. 2. Ребра приведены в порядке присоединения. Сводная таблица переменных представлена в табл. 3. С целью определения «ключевых» переменных набора данных переменные были отсортированы по убыванию значений оптимизационной функции.

Таблица 2

Сводная таблица ребер графа для российских регионов, 1994 год

Вершина 1	Вершина 2	Коэффициент корреляции	Логарифм правдоподобия
retail4	expens4	0,95	170,9
grp4	ind4	0,90	124,6
grp4	inv4	0,84	92,2
assets4	grp4	0,84	91,6
grp4	expens4	0,82	85,2
inc2min4	poor4	–0,81	82,4
commun3	dem3	–0,81	80,5
old4	netw4	0,79	73,4
murders4	expect4	–0,78	72,7
crime4	expect4	–0,73	58,1
retail4	urban4	0,71	52,3
expens4	commun3	–0,69	49,1
expect4	commun3	0,68	46,9
old4	migrat4	0,68	46,3
expens4	inc2min4	0,65	42,4
tempjan	netw4	0,65	42,3
expens4	smentr4	0,61	35,0
grp4	tempjan	–0,60	34,2
periph	netw4	–0,54	26,7
research4	urban4	0,54	25,7
agr4	urban4	–0,52	24,3
native	commun3	0,50	22,2
expens4	avto4	0,48	20,2
house4	unempl4	–0,46	18,4
murders4	infmort4	0,44	16,3

Окончание табл. 2

Вершина 1	Вершина 2	Коэффициент корреляции	Логарифм правдоподобия
research4	netw4	0,45	29,4
murders4	tempjan	–0,57	18,6
ind4	expens4	0,63	15,2
ind4	urban4	0,63	15,3
research4	infmort4	–0,41	15,2
inc2min4	netw4	0,24	15,4
inv4	old4	–0,50	18,2
retail4	smentr4	0,47	13,8

Таблица 3

Сводная таблица переменных для российских регионов, 1994 год

Переменная	Степень	Логарифм правдоподобия	Комбинация	R^2	Качество представления
expens4	7	418	846	0,96	0,88
grp4	5	426	733	0,96	0,84
netw4	5	187	493	0,74	0,72
commun3	4	199	443	0,80	0,82
retail4	3	237	420	0,92	0,85
urban4	4	118	362	0,77	0,74
expect4	3	178	361	0,77	0,74
ind4	3	155	338	0,87	0,69
inc2min4	3	140	324	0,77	0,65
old4	3	138	321	0,73	0,67
murders4	3	108	291	0,66	0,61
tempjan	3	95	278	0,63	0,68
research4	3	70	254	0,45	0,63
inv4	2	110	233	0,76	0,79
smentr4	2	49	171	0,47	0,59
infmort4	2	32	154	0,29	0,59
assets4	1	92	153	0,70	0,72
poor4	1	82	144	0,66	0,54
dem3	1	80	142	0,65	0,74
crime4	1	58	119	0,53	0,69
migrat4	1	46	107	0,46	0,57
periph	1	27	88	0,30	0,49
agr4	1	24	85	0,27	0,21
native	1	22	83	0,25	0,59
avto4	1	20	81	0,23	0,59
house4	1	18	79	0,21	0,21
unempl4	1	18	79	0,21	0,15
student4	0	0	0	0	0,43
patr3	0	0	0	0	0,19

Графовые модели для 1994 и 1999 года приведены соответственно на рис. 6 и 7. В целях экономии места здесь отражена только наиболее существенная часть информации, послужившая основой анализа.

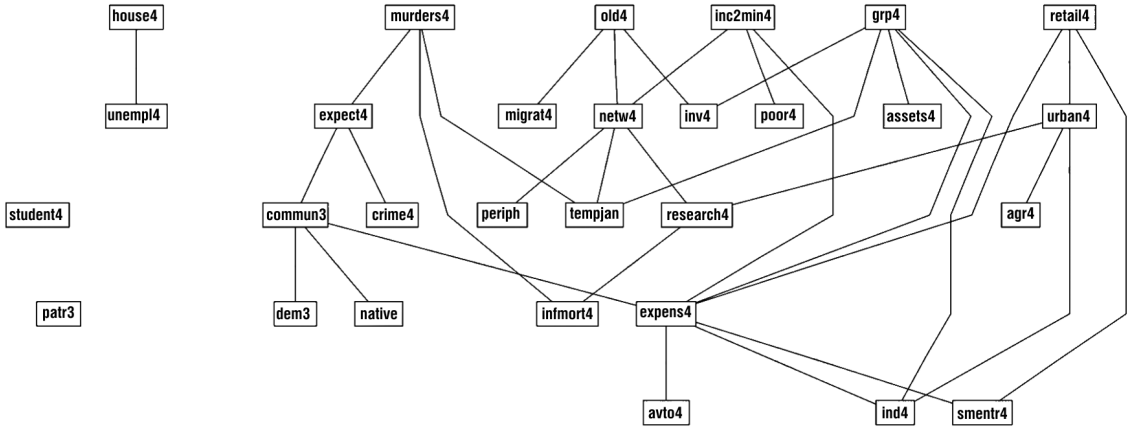

Рис. 6. Графическая модель для российских регионов, 1994 год

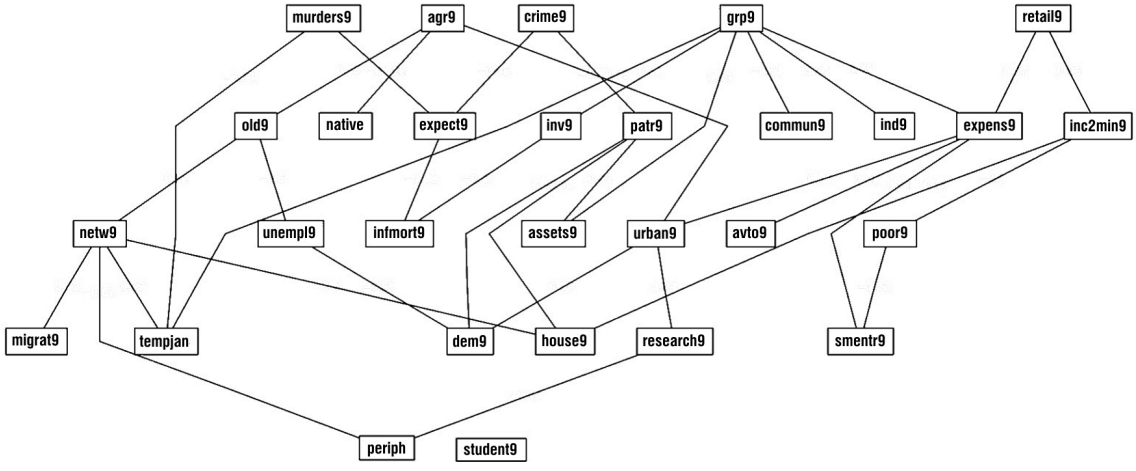

Рис. 7. Графическая модель для российских регионов, 1999 год

В 1994 году ключевыми переменными являлись ВРП, расходы домашних хозяйств, плотность дорожной сети, число научных работников, процент городского населения, «прокоммунистическое» голосование, отношение дохода к прожиточному минимуму, число убийств и процент населения старше трудоспособного возраста (см. рис. 6). В 1999 году структура переменных мало изменилась (см. рис. 7). Переменная patr9 (голосование за «патриотические», правонационалистические партии) отражает здесь усиление протестных настроений в обществе после кризиса 1998 года.

5.2. Введение в анализ структуры переменных

Прежде всего, отметим, что отношение логарифма правдоподобия, объясняемого моделью, и логарифма правдоподобия полного набора данных, возросло с 58% в 1994 году до

67% в 1999 году. Общее качество интерпретации выросло менее существенно: с 0,64 в 1994 году до 0,67 в 1999 году. Это указывает на большую связность наборов данных, поскольку переменные более коррелированны между собой в 1999 году, чем в 1994 году, что также объясняется снижением влияния неэкономических, «принудительных» решений советского периода развития, а также возросшей взаимозависимостью между социальными и экономическими факторами.

Напомним, в чем заключается основная идея метода прямого выбора в графовых моделях (см. подраздел 2.2, I), которую, в частности, реализует алгоритм Демпстера: модель строится путем добавления ребер, дающих наибольшее количество новой информации, измеренной логарифмическим правдоподобием всего набора данных. Процедура добавления новых ребер прерывается согласно особому «правилу остановки». Это правило проверяет гипотезу о том, что коэффициенты частичной корреляции нового добавляемого ребра существенно отличаются от нуля.

Анализ основывается на коэффициентах частичной корреляции. Частичная корреляция между двумя переменными означает наличие корреляции только между этими двумя переменными, влияние всех остальных переменных набора данных не учитывается. Следовательно, если переменная *v*1 в графических моделях связана с переменной *v*3 только через переменную *v*2, считаем, что переменные *v*1 и *v*3 влияют друг на друга только через переменную *v*2.

В процессе анализа предстоит выполнить четыре задачи:

- выбрать ключевые переменные;
- проверить стабильность структуры переменных;
- наблюдать изменения в структуре в течение переходного периода;
- определить точки возможного управляющего воздействия.

5.3. Описание ключевых переменных

ВРП (валовой региональный продукт) на душу населения (grp). На протяжении всего периода 1994–1999 годов эта переменная обнаруживает устойчивые связи с такими переменными, как:

- объем инвестиций на душу населения;
- основные фонды на душу населения;
- расходы домашних хозяйств на душу населения;
- объем промышленного производства на душу населения;
- температура января.

В 1995 году ВРП на душу населения также связан с безработицей, а в 1998–1999 годах — с «прокоммунистическим» голосованием на парламентских выборах 1999 года.

ВРП на душу населения — один из ключевых экономических индикаторов для каждого региона. В то же время в России из-за влияния неравномерности уровня цен[7] и так называе-

[7] Например, в некоторых сибирских регионах из-за необходимости импортировать практически все продукты питания и почти все потребительские товары уровень цен почти в 2 раза выше, чем в центральной части России.

мых «северных надбавок»[8] он имеет значительную географическую составляющую, что подтверждается связями с другими переменными, зависящими от уровня цен, и с температурой января. В российских условиях ВРП практически полностью определяется промышленным производством и наличием природных ресурсов (переменные «основные фонды» и «инвестиции»). Связь с безработицей в 1995 году объясняется имевшим в то время место общим кризисом в российской промышленности и, соответственно, вре́менной остановкой многих предприятий.

Расходы домашних хозяйств на душу населения (expens) связаны с такими переменными, как:

- «прокоммунистическое» голосование;
- отношение дохода к прожиточному минимуму;
- валовой региональный продукт на душу населения;
- объем розничного товарооборота на душу населения;
- объем промышленного производства на душу населения;
- процент городского населения (через объем розничного товарооборота — retail);
- число автомобилей на душу населения.

Эти многочисленные связи сохраняются даже в 1997 году, когда к списку переменных добавились «процент голосов, отданных за демократические партии» и «процент коренного населения». Связь с процентом городского населения, т. е. степенью урбанизации, непрямая, а проходит через объем розничного товарооборота. В целом наш анализ основан на прямых связях, но вследствие некоторых особенностей расчетов, проводимых Госкомстатом, «объем розничного товарооборота» (retail) и «расходы домашних хозяйств на душу населения» (expens) оказались слишком сильно коррелированными (0,95), и в данном случае без ущерба для смысла общее правило можно нарушить. Также представляется возможным оставить одну из этих переменных за рамками нашего анализа.

Переменная «расходы домашних хозяйств на душу населения» является агрегированным индикатором экономической деятельности и поэтому имеет положительную корреляцию со «степенью урбанизации» и отрицательную — с «процентной долей голосов, отданных за прокоммунистические партии».

Проанализировав табл. 2 и 3, приходим к убеждению, что высокие значения логарифмов правдоподобия первых двух добавленных связей — между объемом розничного товарооборота и расходами домашних хозяйств на душу населения, а также между ВРП и объемом промышленного производства — в значительной мере определяют высокие значения логарифмов правдоподобия этих экономических индикаторов. Особенно очевидным это становится в случае с объемом промышленного производства, чье высокое значение логарифма объясняется его связью с ВРП.

Эти две первые переменные — ВРП и «расходы домашних хозяйств на душу населения» — имеют хорошее качество представления и высокое значение R^2. Последнее означает, что их значения легко оценить исходя из смежных переменных.

[8] Повышающие коэффициенты, установленные государством в районах Крайнего Севера и приравненных к ним районов (64% территории страны, но менее 6% населения). Применяются к зарплатам и пенсиям для компенсации тяжелых климатических условий и высокого уровня цен.

Отношение дохода к прожиточному минимуму (inc2min) является агрегированным индикатором благосостояния, что подтверждается стабильностью его связей с такими переменными, как:

- процентная доля населения, живущего ниже прожиточного минимума (бедного населения);
- плотность транспортной сети и, следовательно, уровень развития инфраструктуры;
- расходы домашних хозяйств или объем розничного товарооборота, поскольку, как уже отмечалось, Госкомстат рассчитывал расходы домашних хозяйств исходя из объема розничного товарооборота.

В 1997 году возникает новая связь — с жилищным строительством. В этот период российская экономика становится более рыночно-ориентированной, и теперь жилищное строительство в большей степени зависит от дохода населения, чем от внеэкономических факторов.

Ожидаемая продолжительность жизни при рождении (expect). Это стабильный агрегированный индикатор качества населения [Айвазян (2002)], суммирующий влияние таких переменных, как «число убийств и преступлений на душу населения», «детская смертность», «плотность транспортной сети», а в 1997 году также «жилищное строительство».

Процент населения старше трудоспособного возраста (old) всегда остается ключевой переменной любого набора данных в российских регионах и в значительной степени противопоставляет давно заселенные регионы недавно заселенным[9].

Плотность транспортной сети (netw). Эта переменная не только отражает уровень развития инфраструктуры региона и его географию[10], но также является агрегированным индикатором качества жизни, что подтверждается его связью с отношением дохода к прожиточному минимуму и, в последние годы, связями с жилищным строительством и миграцией.

Численность населения, живущего ниже прожиточного минимума (poor). Значение этой переменной возросло в 1999 году. Она имеет связи не только с отношением дохода к прожиточному минимуму, но и с числом малых предприятий.

Остальные переменные, сохранившие свое значение на протяжении всего периода исследования, — это «инвестиции» (inv), «число убийств и покушений на убийство» (murders) и «число малых предприятий» (smentr). Они отражают соответственно эффект от инвестиций как таковых, эффективность государственного управления в сфере обеспечения безопасности населения, а также уровень предпринимательской деятельности в регионе.

5.4. Анализ структуры «поля переменных»; стабильность и изменения

Как уже было отмечено, в течение переходного периода наблюдалось усиление связности набора данных и взаимного влияния между социальными, географическими и экономическими индикаторами.

[9] Давно заселенные европейские регионы характеризовались наивысшим процентом населения старших возрастов. Многие молодые люди мигрировали оттуда в недавно заселенные регионы (некоторые регионы Сибири и Дальнего Востока) с крайне низким процентом населения старших возрастов.

[10] Вследствие административных решений в советский период транспортные расходы при перевозках между заводами не оказывали влияния на цены товаров.

В целом можно говорить о своего рода «поле переменных» — мы начинаем с *переменных, характеризующих экономическую и предпринимательскую деятельность, качество населения*[11] (число малых предприятий, процентная доля населения, живущего ниже прожиточного минимума, отношение дохода к прожиточному минимуму и объем розничного товарооборота). Далее переходим к *общим макроэкономическим индикаторам*, к которым относятся ВРП, объем промышленного производства, объем инвестиций, основные фонды, а затем через *инфраструктурные* и *географические индикаторы* (степень урбанизации, густота дорожной сети, детская смертность, число убийств и покушений на убийство и др.) — к *социальным индикаторам*, таким как уровень безработицы, процент лиц старше трудоспособного возраста и, наконец, миграция.

Экономические переменные характеризуются стабильной структурой корреляций. Связи между географическими и социальными индикаторами больше изменяются во времени, отражая новые тенденции в данных. Например, если в 1994 году переменная «жилищное строительство» была практически изолирована от остальных данных, то в 1999 году она оказалась связанной с такими переменными, как «протестное "патриотическое" голосование», «плотность транспортной сети» и «отношение дохода к прожиточному минимуму».

Наиболее нестабильны политические переменные: они перемещаются по графику после каждых новых выборов, которые нередко полностью меняют «смысл» этих переменных. В разные периоды времени голосование за одни и те же партии может отражать различные процессы и тенденции, существующие в обществе. Например, переменная, отражающая «патриотическое» голосование, т. е. протестные правонационалистические настроения, в 1994 году была изолированной. В 1999 году эта же переменная имела четыре связи: с продемократическим голосованием, жилищным строительством, основными фондами и уровнем преступности, что стало политическим следствием финансового кризиса 1998 года, когда большинство развитых регионов с многочисленным средним классом оказалось в числе наиболее пострадавших от обвального падения рубля. Сильная связь с уровнем преступности также является неотъемлемым свойством переменной «патриотического» голосования[12].

Политические переменные обнаруживают стабильную связь с процентом (долей) городского населения, что, как известно из российской политической практики, позволяет предсказывать электоральные предпочтения в регионах. Этот эффект был отмечен специалистами в области экономической географии[13]. В целом процент городского населения в значительной степени определяет как социально-экономический облик региона, так и пути его развития.

Переменная, отражающая численность студентов, остается изолированной в течение всего изучаемого периода. Возможно, мы должны обратить на нее особое внимание: либо она некорректно рассчитывается, либо научно-исследовательская деятельность и пред-

[11] *Качество населения* — синтетический показатель, отражающий воспроизводство, демографическую структуру и физическое здоровье населения, способность образовывать и сохранять семьи, уровень образования и культуры, уровень квалификации населения.

[12] Заключенные в России голосуют за правые националистические партии. Тюрьмы, как правило, расположены в нескольких северных районах, например в Республике Мордовия. Вышедшие на свободу часто вынуждены селиться неподалеку от своих бывших тюрем, поскольку согласно некоторым законам, оставшимся в силе с советских времен, они лишаются прописки в прежних местах проживания. Среди этого контингента населения также весьма высок уровень рецидивной преступности.

[13] Личные беседы автора с проф. Л. Смирнягиным, канд. геогр. наук Е. Скатерщиковой и др.

принимательская деятельность не имеют связи с университетами и институтами, что может означать наличие проблем в системе высшего образования России (см. Приложение 3.3).

Переменные, относящиеся к одной группе, можно обнаружить в разных частях «поля». Например, такие демографические переменные, как «процент городского населения» и «миграция» расположены очень далеко друг от друга. Первая из них отражает экономическую деятельность, а вторая — в основном географические аспекты.

Такие переменные, как «объем жилищного строительства», «уровень безработицы», переменные, отражающие социальное неравенство («процент бедного населения» и «отношение среднего дохода к прожиточному минимуму»), также перемещались по «полю». Это перемещение представляется важным, поскольку оно отражает изменения, которые претерпело общество за годы перестройки. В 1994 году эти переменные были либо практически полностью изолированными, либо более связанными с географическими переменными. В 1999 году положение этих переменных (за исключением безработицы) определялось экономической деятельностью в регионе.

5.5. Точки возможного управляющего воздействия

Понимание структуры переменных позволяет выяснить, какие внешние воздействия могли бы привести к наиболее ощутимым результатам. Например, вызывает сомнение, способна ли такая изолированная переменная, как «численность студентов», повлиять на социально-экономическое положение в регионе. Представляется, что самыми логичными точками приложения управляющего воздействия могут быть: *промышленное производство*, увеличение которого подразумевает рост ВРП и снижение процентной доли населения, живущего ниже прожиточного минимума, и *детская смертность*.

Также можно предположить, что возможность приобретать в кредит дома, квартиры и автомобили окажет положительное влияние на уровень расходов домашних хозяйств и отношение дохода к прожиточному минимуму. Высокий уровень расходов домашних хозяйств, в свою очередь, имеет высокую положительную корреляцию с отношением дохода к прожиточному минимуму. Оживление предпринимательской деятельности (увеличение числа малых предприятий на душу населения) — еще один способ повысить расходы домашних хозяйств на душу населения и снизить процентную долю населения, живущего ниже прожиточного минимума (бедного населения).

Помимо этого можно рассмотреть *жилищное строительство* вместе с *плотностью транспортной сети* как возможные экзогенные переменные, или «точки влияния» для стимулирования развития регионов. Поэтому государственные инвестиции в этих двух направлениях могли бы дать существенный положительный результат.

Тем не менее отметим, что для подтверждения изложенных гипотез нам необходимо провести тесты Грейнджера, чтобы проверить, какие переменные являются «источниками», а какие — «восприемниками» того или иного влияния.

5.6. Анализ групп регионов (сравнение структур регионов европейской части России и Сибири)

Сравним теперь структуру переменных по группам регионов. Низкое значение отношения числа наблюдений к числу переменных сокращает объем информации по каждой группе, и при разделении более чем на две группы последние в нашем случае окажутся слишком маленькими для того, чтобы обеспечить получение значимых результатов.

При разделении регионов на две группы применялись различные критерии. Сначала были использованы *рейтинги* [Айвазян (2002)], полученные в соответствии со следующими синтетическими критериями:

- качество населения;
- жизненный уровень;
- качество социальной сферы.

Далее мы перешли к разделению по *ключевым переменным набора данных,* к числу которых относятся «отношение дохода к прожиточному минимуму», «ожидаемая продолжительность жизни при рождении», «число убийств на душу населения» и др.

Графические модели этих разделений сохраняют структуру переменных почти без изменений. Происходит только незначительное снижение числа связей за счет уменьшения отношения числа наблюдений к числу переменных. Некоторые переменные стали более важны в отдельных наборах данных: например, «"прокоммунистическое" голосование» в 1994 году или «объем сельскохозяйственного производства» в 1999 году. Причиной этому, возможно, послужили: в 1994 году — политическая нестабильность; в 1999-м — возрождение сельского хозяйства после падения рубля в 1998 году и резкого снижения объема импорта.

В данной статье представлены результаты только одного, самого интересного, «*естественного*» *разделения* — на регионы европейской части страны и регионы Сибири в период окончания перестройки (1999 год). Соответствующие графические модели для 1999 года приведены на рис. 8 и 9. Уральские горы служат традиционной границей между европейской и азиатской (сибирской) частями России, собственно же уральские регионы оказались разделенными. Мы сочли, что республики Башкортостан и Удмуртия, Оренбургская и Пермская области относятся к европейской части, тогда как Курганская, Челябинская и Свердловская области — к Сибири.

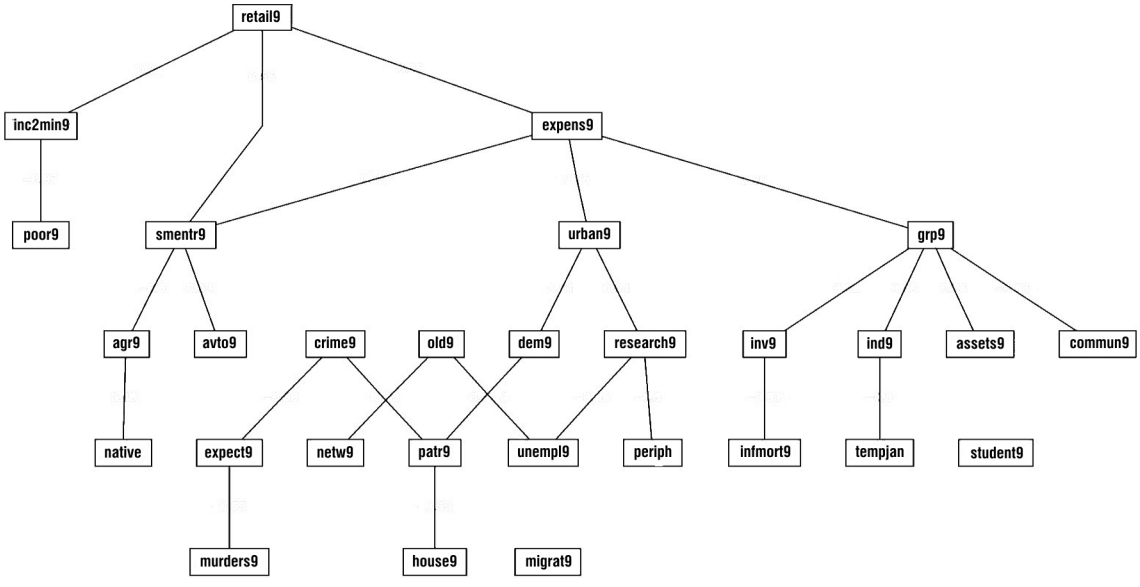

Рис. 8. Графическая модель для российских регионов европейской части страны, 1999 год

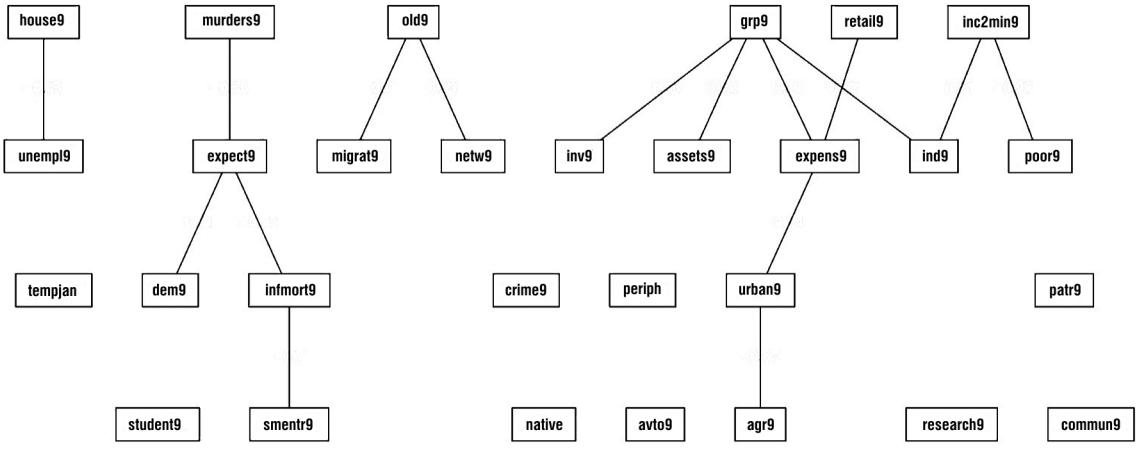

Рис. 9. Графическая модель для российских регионов сибирской (азиатской) части страны, 1999 год

Графическая модель европейских регионов характеризуется растущей ролью переменной, отражающей число малых предприятий на душу населения. Эта переменная связана с расходами домашних хозяйств, объемом розничного товарооборота, объемом сельскохозяйственного производства и числом автомобилей (все вышеперечисленные — на душу населения). Таким образом, наряду с расходами домашних хозяйств и ВРП на душу населения число малых предприятий становится ключевым экономическим индикатором.

Вследствие малочисленности сибирских регионов их граф не содержит большого числа связей. Два экономических индикатора — ВРП и расходы домашних хозяйств на душу населения, а также такой индикатор качества населения, как ожидаемая продолжительность жизни при рождении, являются ключевыми переменными этого набора данных.

Сравнивая графы обоих наборов данных, можно сделать вывод, что наличие природных ресурсов и тяжелой промышленности по-прежнему является определяющим фактором экономической деятельности за Уралом, в то время как экономическая деятельность в европейской части России становится все более связанной с рыночными формами производства, в частности с числом малых и средних предприятий.

5.7. Некоторые выводы по структуре переменных

Обобщая анализ, можно отметить:

1. Мы увидели усилившуюся «связность» факторов, определяющих социально-экономическое положение и развитие российских регионов. В 1994 году внеэкономические факторы, унаследованные от советского периода развития, обусловили ряд необычных тенденций в данных. Например, прямую связь между числом научных работников и густотой дорожной сети, между процентом уроженцев региона от общей численности его населения и числом голосов, поданных за коммунистов, а также между процентом населения старше трудоспособного возраста и миграцией. Как видно на графе 1999 года, эти и другие особенности структуры переменных исчезли за годы переходного периода, тогда как число тех свя-

зей между переменными, которые объясняются традиционными социально-экономическими теориями, оставалось сравнительно стабильным или даже увеличилось.

2. Нам представляется, что главный результат этого раздела состоит в обнаружении «поля переменных», в котором позиции основных социально-экономических, географических и демографических переменных оставались неизменными, а политических (в нашем случае электоральных) — перемещались.

6. Выводы и дальнейшие направления исследования

В работе дано краткое введение в графовые модели и подробно представлен модифицированный алгоритм Демпстера, а также технология его применения для различных наборов данных. Кроме того, описано его применение к сравнительному исследованию российских регионов во второй половине периода перестройки (1994–1999 годы).

Распространение настоящего исследования на другие периоды времени, больший набор переменных или другие страны, а также сопоставление полученных результатов с результатами настоящего исследования может обеспечить проверку нашей гипотезы о существовании в социально-экономических исследованиях некой общей структуры переменных. Отклонение от такой структуры может свидетельствовать об экономическом и социальном своеобразии страны или региона в течение определенного периода времени.

В настоящее время автор в сотрудничестве с Татьяной Рыбниковой (ЦЭМИ) и Жераром Антилем (Женевский университет) продолжает эту работу соответственно для базы данных регионов России за 1997–2007 годы [Рыбникова, Вайнберг Аллен (2008)] и базы данных Лозаннского Института Развития, публикуемой в «Ежегоднике мировой конкурентоспособности» [World Competitiveness Yearbook (2008)]. Наличие институциональных переменных в обеих совокупностях данных позволяет проверить гораздо более широкий и интересный набор гипотез по структуре (полю) переменных.

Результаты этих работ мы хотели бы опубликовать в последующих номерах журнала.

Приложение 1

Алгоритм поиска кратчайшей траектории Дейкстры (Dijkstra)

Алгоритм для поиска к р а т ч а й ш е й траектории из данной вершины i ко всем другим вершинам разработан Дейкстрой [Dijkstra (1959)]. Сложность этого алгоритма равна $O(p^2)$ [Cormen et al. (1990)].

Определения

$G(X, U)$ — граф;

p — число вершин;

$l(i, j)$ — длина ребра $(i, j) \in U$;

Γ_i — множество вершин, смежных с вершиной i;

$\Pi^*(i)$ — длина кратчайшей траектории из вершины 1 к вершине i;

$\Pi^*(i) = 0$.

Введение

- Выполняем $p - 1$ итерацию. В начале каждой итерации имеем два множества: множество S всех уже исследованных вершин и множество $\overline{S} = X \setminus S$ неисследованных вершин. На первой итерации $S = \{1\}$.
- Каждая вершина имеет метку $\Pi(i)$ со свойством:

$$\Pi(i) = \begin{cases} \Pi^*(i), & \text{если } i \in S, \\ \min\limits_{k \in \bar{S} \cap \Gamma_i} \{\Pi(k) + l(k, i)\}, & \text{если } i \in \bar{S}. \end{cases}$$

• Значение $\Pi(i)$ для $i \in \bar{S}$ соответствует кратчайшей траектории из вершины 1 к вершине i, когда все вершины, за исключением вершины i, принадлежат S.

Алгоритм 3. Поиск кратчайшей траектории Дейкстры

1. $\bar{S} = X \setminus \{1\}, \Pi(1) = 0$
2. $\Pi(i) = l(1, i)$, если $i \in \Gamma_1$, и ∞ в противном случае
3. **while** $\bar{S} \neq \varnothing$ **do**
4. Выбрать $j \in \bar{S}$ так, чтобы $\Pi(j) = \min\limits_{i \in \bar{S}} \Pi(i)$
5. $\bar{S} = \bar{S} \setminus \{j\}$
6. **for** $i \in \Gamma_j \cap \bar{S}$ **do**
7. $\Pi(i) = \Pi(j) + l(i, j)$
8. **end for**
9. **end while**

Поскольку граф нециклический, то существует единственная траектория от вершины 1 к каждой другой вершине i. Таким образом, оператор 7 из алгоритма 3 упрощается. В исходном алгоритме он имеет вид

$$\Pi(i) = \min \{\Pi(i),\ \Pi(j) + l(i, j)\}.$$

Приложение 2

Список переменных за 1994–1999 годы

agr4-9 — сельскохозяйственное производство на душу населения, в текущих ценах.

assets4-9 — основные фонды на душу населения, в текущих ценах.

avto4-9 — число автомобилей на 1000 человек.

commun3, 5, 9 — голоса, поданные за коммунистов и их союзников на выборах 1993, 1995 и 1999 годов, в процентах.

crime4-9 — уровень преступности: число преступлений на 1000 человек.

dem3, 5, 9 — голоса, поданные за демократические партии и их союзников на выборах 1993, 1995 и 1999 годов, в процентах.

expect4-9 — ожидаемая продолжительность жизни при рождении, в годах.

expens4-9 — расходы домашних хозяйств на душу населения, в текущих ценах.

grp4-9 — валовой региональный продукт (ВРП) на душу населения, в текущих ценах.

house4-9 — жилищное строительство на 1000 человек, в м^2.

inc2min4-9 — отношение среднего дохода в денежном выражении к прожиточному минимуму в денежном выражении, в процентах.

ind4-9 — промышленное производство на душу населения, в текущих ценах.

infmort4-9 — уровень детской смертности.

inv4-9 — инвестиции в основные фонды на душу населения, в текущих ценах.

migrat4-9 — миграция на 1000 человек.

murders4-9 — число убийств и покушений на убийство на 1000 человек.

native — доля населения родившегося в регионе, в процентах.

netw4-9 — плотность транспортной сети, в км/м^2.

old4-9 — лица старше трудоспособного возраста, процент от общей численности населения.

patr3, 5, 9 — голоса за «патриотические» (правые националистические) партии и их союзников на выборах 1993, 1995 и 1999 годов, в процентах.

periph — периферийность, в баллах.

poor4-9 — численность населения с доходами ниже прожиточного минимума (так называемого бедного населения), в процентах.

research4-9 — число научных работников на 1000 человек.

retail4-9 — розничный товарооборот на душу населения, в текущих ценах.

smentr5-9 — число малых предприятий на 1000 человек.

student4-9 — число студентов на 1000 человек.

tempjan — температура января, в градусах.

unempl4-9 — число безработных на конец года, на душу населения.

urban4-9 — городское население, в процентах.

Приложение 3

К вопросу о подготовке данных

Как уже отмечалось, мы проводили анализ на основе 77 многомерных наблюдений, представлявших регионы (субъекты Федерации). Автономные округа, а также республики Ингушетия и Чечня были исключены из исследования.

Набор данных покрывает шестилетний период с 1994 по 1999 год и содержит 29 переменных. Почти все переменные представляют собой официальные данные Госкомстата. К сожалению, наш набор не включает переменные, относящиеся к внешней торговле, поскольку эти данные по регионам имеются в наличии лишь с 1998 года.

В нашем распоряжении не было также комплексного индикатора климатических условий российских регионов. Вместо него была использована средняя температура января, причем с тем, чтобы зафиксировать этот индикатор, пришлось использовать одну и ту же температуру для всех лет на протяжении изучаемого периода, в нашем случае температуру января 1997 года. Переменная «процент населения, родившегося в регионе» приведена по данным переписи населения 1988 года. Плотность транспортной сети рассчитывалась как среднее значение плотности автодорожной и железнодорожной сетей (число километров дорог, деленное на площадь территории).

Мы не располагали переменными, описывающими институциональное развитие, а также действия правительства и проведение реформ, поэтому были вынуждены использовать результаты выборов при изучении политико-институциональных характеристик регионов.

Мы использовали переменные в текущих ценах, поскольку в постоянных ценах многие из них недоступны. В частности, ВРП на душу населения в постоянных ценах Госкомст России регистрировал только начиная с 1996 года. Тем не менее, если принять рабочую гипотезу об одинаковой инфляции во всех регионах[14], использование переменных в текущих ценах не влияет на результаты, так как мы всегда работаем с нормированными переменными.

[14] Вообще говоря, такая гипотеза является достаточно грубой.

Также следует отметить, что 1994 год стал первым годом расчета ВРП, и специалисты, которые им занимались, еще только знакомились с методологией. Помимо этого существует проблема расчета основных фондов. Предшествующие 1990–1993 годы были временем высокой инфляции, и в 1994 году были пересчитаны еще не все основные фонды[15]. В результате основные фонды за 1994 год превышают ВРП только в 3 раза, тогда как обычно это соотношение равняется 10. Есть основания предположить, что эти проблемы относились не только к центральному управлению Госкомстата, но и к его региональным отделениям, поэтому их данные не должны приводить к существенным ошибкам при региональных сравнениях за один и тот же период.

3.1. Переменная «периферийность» (periph)

«Периферийность» понимается нами как удаленность от границ и основных экономических и культурных центров страны. Данная переменная построена автором экспертно на основе подробного справочника по российским регионам [Europa Publications Limited (1999)]. Были использованы три критерия классификации:

I. Близость (в географическом и транспортном смысле) к центрам страны (международным аэропортам);

II. Близость к границам и морю, особенно наличие портов.

III. Наличие городов с населением более 500 тыс. человек.

Были определены пять групп и пятнадцать подгрупп:

1. Центры страны, прибыльные порты и «полезные» границы.

1.1. Москва, Санкт-Петербург.

1.2. Приморский край, Архангельская, Калининградская, Ленинградская, Нижегородская, Новгородская и Псковская области.

1.3. Республика Карелия, Хабаровский и Краснодарский края, Калужская, Мурманская, Ростовская и Тульская области.

2. Центр европейской части России и менее прибыльные порты.

2.1. Республики Дагестан и Татарстан; Амурская, Кировская, Пермская, Самарская, Сахалинская, Саратовская, Ульяновская, Волгоградская, Вологодская, Ярославская и Тверская области.

2.2. Республики Чувашия и Марий Эл, Астраханская область.

2.3. Республики Башкирия, Северная Осетия, Мордовия и Удмуртия, Орловская и Пензенская области.

3. Внутренние районы европейской части России, Уральские горы и часть Сибири.

3.1. Ивановская, Костромская, Челябинская, Курганская, Рязанская, Ставропольская, Тамбовская и Владимирская области.

3.2. Республики Калмыкия, Карачаево-Черкесия и Кабардино-Балкария, Белгородская, Брянская, Курская, Липецкая, Смоленская и Воронежская области, Ставропольский край.

3.3. Новосибирская, Омская, Томская и Тюменская области.

[15] Частная информация от канд. экон. наук Зайцевой.

4. Середина сибирской и север европейской части России.

4.1. Кемеровская, Челябинская, Новосибирская и Оренбургская области.
4.2. Республика Коми, Красноярский край, Иркутская область.
4.3. Архангельский край.

5. Удаленные регионы: вдали от центра, нет удобного доступа за границу.

5.1. Республики Адыгея, Саха-Якутия, Бурятия, Читинская, Камчатская области, Хабаровский край.
5.2. Еврейская автономная и Магаданская области.
5.3. Республики Алтай и Тыва.

Для «квантификации» этой классификации применяли схему Морриса и Адельмана [Morris, Adelman (1988)]. Мы использовали линейную шкалу от 0 до 100. Максимальное значение 90 соответствует группе 1, минимальное 10 — группе 5, значения для остальных групп были равномерно распределены между ними. Значения для подгрупп были определены таким образом, чтобы расстояние между последней подгруппой одной группы и первой подгруппой последующей группы было примерно в 2 раза больше, чем расстояние между подгруппами внутри группы. В частности, 95 соответствует подгруппе 1.1 (Москва и Санкт-Петербург), 90 — подгруппе 1.2, 85 — подгруппе 1.3, а 75 — подгруппе 2.1. И наконец, мы умножаем все значения на –1, чтобы переменная «периферийность» принимала свое максимальное значение для наиболее «удаленных» регионов.

3.2. Отсутствующие переменные

Сведения о количестве малых предприятий за 1994 год недоступны, поэтому в 1994 году мы воспользовались переменной 1995 года.

В трех случаях отсутствуют данные для Москвы и Санкт-Петербурга:

1. Температура января (tempjan). Оба отсутствующих значения берутся равными соответствующим значениям для Московской и Ленинградской областей.
2. Плотность транспортной сети (netw). Значения для городов Москва и Санкт-Петербург взяты из данных для Московской и Ленинградской областей соответственно.
3. Данные по сельскохозяйственной продукции (agr). Они берутся равными нулю.

Другие отсутствующие значения[16] были рассчитаны с помощью классических методов регрессии, изложенных в работе Шафера [Schafer (1996)]. Предположим, что отсутствует значение переменной X для года t_0 для региона r_0. Мы исключаем регион r_0, а для всех остальных переменных строим модель регрессии, считая X_{t_0} зависимой переменной. В качестве объясняющих используем переменные X_{t_i} $(i = 1, ..., T)$ для T других лет и l других различных переменных $Y_{t_0, j}$ $(j = 1, ..., l)$ для того же года t_0. Отсюда получаем

$$X_{t_0} = f(X_{t_1}, ..., X_{t_T}, Y_{t_0,1}, ..., Y_{t_0,l}).$$

Затем полученное новое соотношение используется для расчета значений для региона r_0.

[16] Всего 12 значений: процент населения с доходами ниже прожиточного минимума в Республике Дагестан в 1994 году и в Еврейской автономной области в 1994–1998 годах; отношение среднего дохода к прожиточному минимуму для Республики Северная Осетия — Алания и Еврейской автономной области в 1994 году; число безработных на душу населения для Северной Осетии — Алании в 1994 году; число студентов на душу населения в Ленинградской области в 1994–1996 годах.

Наилучшая модель выбирается путем использования C_p-статистики Мэллоуза [Mallows (1973)]. Данный метод вносит смещение. Этого можно избежать, например, прибавив к оцененному результату нормально распределенную случайную ошибку. Тем не менее в нашем случае вследствие пренебрежимо малого числа отсутствующих значений (12) по отношению к общему объему данных (более 10 тысяч наблюдений) этим смещением можно пренебречь.

3.3. Проверка на нормальность и преобразование переменных

Перед применением алгоритма Демпстера или модифицированного алгоритма Демпстера построения графовых моделей необходимо удостовериться, что эмпирическое распределение наших данных близко к нормальному. Мы делаем это с помощью «квантиль—квантиль» (К–К) графиков нормального распределения, представленных на рис. 10.

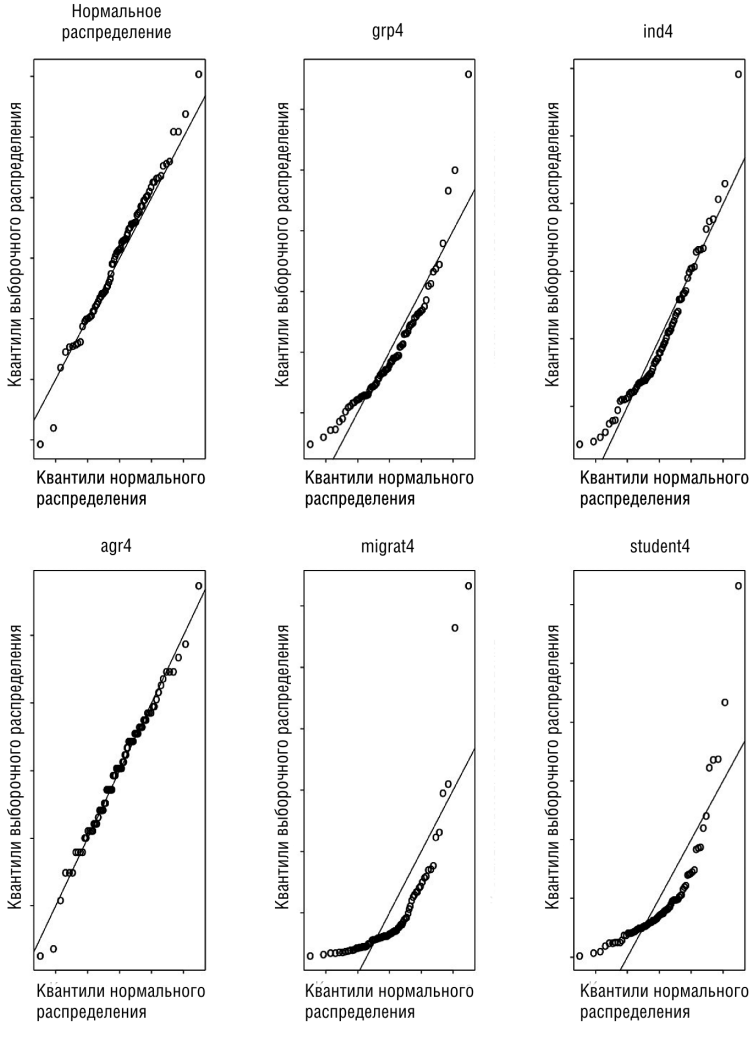

Рис. 10. Графики «квантиль–квантиль» (К–К) нормального распределения для некоторых переменных 1994 года

Сортируем значения переменной в порядке возрастания, и для каждого значения откладываем по горизонтали квантили нормального распределения, а по вертикали — квантили выборочного распределения. Соответственно, значения нормально распределенной переменной должны лежать строго на прямой линии.

Первый из шести графиков рисунка представляет распределение выборочной нормальной переменной и представлен здесь для «калибровки глаза» на относительную важность отклонений от прямой линии, а значит, как мы уже отметили, от нормальности [Welsh (1996)]. Остальные графики представляют пять переменных 1994 года: ВРП на душу населения (grp4), промышленное производство на душу населения (ind4), сельскохозяйственное производство на душу населения (agr4), миграция на 1000 человек (migrat4) и количество студентов на душу населения (student4).

Мы видим, что за исключением переменной сельскохозяйственного производства на душу населения (agr), которую можно считать приблизительно нормальной, распределения оставшихся четырех переменных далеки от нормального, что обусловливает необходимость их преобразования [Welsh (1996), Хальд (1956)].

Список преобразований переменных представлен в табл. 4. Все указанные преобразования дают удовлетворительные результаты, кроме преобразования переменной числа студентов на душу населения (student). Это единственная переменная набора данных, к которой мы не можем подобрать правильное преобразование, возможно, в дальнейшем к ней надо будет применить метод «нормальных меток» Ван-дер-Вардена (нормальное распределение рангов переменных) [Ван-дер-Варден (1960), Благовещенский (2008)].

Таблица 4

Преобразование переменных

Преобразование	Применяется к переменным
Логарифм	assets, grp, inv, retail, expense, inc2min, research, crime, murders, poor, unempl, student, infmort, dem
Корень	ind, house, commun, netw
Экспонента стандартизированных данных	native, migrat
Не преобразовываются	agr, avto, urban, old, expect, patr, smentr, tempjan, periph

Список литературы

Айвазян С. А. Анализ категорий качества жизни населения субъектов Российской Федерации: их измерение, динамика, основные тенденции (по статистическим данным за 1997–1999 гг.) // *Уровень жизни населения России*. 2002. № 11.

Благовещенский Ю. Н. Тайны корреляционных связей в статистике. М.: Научная книга, 2008.

Ван-дер-Варден Б. Л. Математическая статистика. М.: Иностранная литература, 1960.

Рыбникова Т. С., Вайнберг Аллен А. Л. Описание базы данных институциональных показателей по регионам РФ 1999–2007 гг. www.cemi.rssi.ru, 2008.

Хальд А. Математическая статистика с техническими приложениями. М.: Иностранная литература, 1956.

Ammor N., Chah Slaoui S. Algorithme de noyaux stables // In *XXXIIe Journées de Statistiques. Actes*. GRESTAF. Fès. Maroc. 2000.

Borgatti S. P. How to explain hierarchical clastering // *Connectons*. 1994. V. 17(2).

Cormen T. H., *Leiberson C. E.*, *Rivest R. L.* Introduction to Algorithms. MIT Press, 1990.

Dempster A. Covariance selection // *Biometrics*. 1972. V. 28. March.

Diday E. Une nouvelle methode en classification automatique et reconnaissance des formes // *Revue de statistique appliqué*. 1971. V. 19(2).

Diday E., *Lemaire J.*, *Pouget J.*, *Testu F.* Eléments d'analyse de données. Paris: Bordas, 1982.

Dijkstra E. W. A note on two problems in connection with graphs // *Numerische Mathematik*. 1959. V. 1.

Europa Publications Limited. The Territories of the Russian Federation. Old Woking, Surrey, UK: The Gresham Press. 1999.

Fienberg S. E. Graphical methods in statistics // *American Statistician*. 1979. V. 33.

Johnson S. Hierarchical clustering schemes // *Psychometrika*. 1967. V. 2.

Morris C. T., *Adelman I.* Comparative Patterns of Economic Development 1850–1914. Baltimore and London: John Hopkins University Press, 1988.

Mallows C. L. Some comment on c_p // *Technometrics*. 1973. V. 15.

Schafer J. Analysis of Incomplete Multivariate Data. London: Chapman & Hall, 1996.

Venables W., *Ripley B.* Modern Applied Statistics with S-Plus. New York: Springer-Verlag, 1994.

Weinberg A. Quantitative analysis of the situation and development of Russian regions during the transition period. Thèse de Doctorat. Geneva: University of Geneva, 2007.

Welsh A. H. Aspects of Statistical Inference. New York: John Wiley & Sons, 1996.

World Competitiveness Yearbook. IMD. Lausanne, 2008.