УДК 004.67

МЕТОД СТРУКТУРНОЙ ГРУППИРОВКИ ОБЪЕКТОВ В ЗАДАЧЕ ПОВЫШЕНИЯ ДОСТОВЕРНОСТИ ОЦЕНОК ПОКАЗАТЕЛЕЙ МОНИТОРИНГА ПО МАЛЫМ ВЫБОРКАМ[1]

А.А. ДОРОФЕЮК, Ю.А. ДОРОФЕЮК, А.Л. ЧЕРНЯВСКИЙ
Институт проблем управления РАН, г. Москва
daa2@mail.ru

Ключевым инструментом анализа и моделирования экономического развития является мониторинг социально-экономических показателей в разрезе субъектов РФ. Главная проблема, с которой сталкиваются статистики развитых стран, – это проблема коррекции статистических данных для малых (нерепрезентативных) выборок, которые обычно возникают из-за недостаточного финансирования выборочных обследований. В настоящей работе предложен новый метод анализа малых выборок, позволяющий получать достаточно точные оценки без уменьшения оперативности мониторинга. Он основан на современной методологии интеллектуального анализа данных, в том числе на алгоритмах структурно-классификационного анализа.

Ключевые слова: *структурно-классификационный анализ, нерепрезентативные выборки, мониторинг динамических объектов.*

[1] Работа выполнена при частичной поддержке РФФИ, проекты 11-07-00178-а, 11-07-00735-а, 13-07-00992.

Проблема недостаточной представительности результатов выборочных статистических обследований (мониторинга) в случае малых, нерепрезентативных выборок особенно остро проявляется при формировании статистических данных в разных структурных разрезах (например, в региональном разрезе, по видам экономической деятельности, по формам собственности, видам продукции, половозрастным группам и т.д.).

Разработаны и широко применяются различные модели и процедуры сглаживания временных рядов, основанные на агрегации данных за несколько временных интервалов [1,2]. В частности, такие процедуры, как *X*12-*ARIMA* (разработчик Бюро Цензов США), *TRAMO-SEATS* (разработчик – Банк Испании), рекомендованы ОЭСР и Евростатом в качестве стандартных методов сезонного сглаживания и применяются на практике многими национальными статистическими органами. Они реализованы в виде специального программного обеспечения *DEMETRA* [3]. Модели типа ARIMA [1] хорошо решают задачу сглаживания временного ряда, но приемлемое качество сглаживания достигается в них лишь в том случае, если для анализа используется достаточно большой отрезок этого ряда. Данные модели весьма инерционны, они не реагируют на резкие изменения показателей, которые происходят, например, во время кризисных ситуаций. Уловить такие изменения способны только самые простые (с точки зрения теории временных рядов) методы с малой памятью, типа метода скользящего среднего [1]. Но эти методы, во-первых, еще более чувствительны к размеру выборки, а во-вторых, являются недостаточно оперативными, так как для получения несмещённых оценок за текущий период времени требуются данные как за предыдущие, так и за будущие периоды.

В работе предложен новый метод повышения достоверности статистических показателей для малых (нерепрезентативных) выборок, позволяющий получать достаточно точные оценки без уменьшения оперативности мониторинга. Он основан на современной методологии интеллектуального анализа сложноорганизованных данных, в том числе на использовании методов структурно-классификационного анализа [4].

1. СОДЕРЖАТЕЛЬНАЯ ПОСТАНОВКА ЗАДАЧИ

Постановка задачи и описание метода даются на примере задачи ежемесячного мониторинга некоторого показателя функционирования социально-экономических объектов. При этом по некоторым причинам (в основном связанным с недостаточностью финансирования) имеющийся объём выборки обеспечивает представительные данные только по системе в целом (обычно это – Российская Федерация) и по некоторым крупным объектам (регионам). Для большинства же объектов достоверно оценить

значения исследуемого показателя непосредственно по выборочным данным не удаётся. В качестве примера на рис. 1 приведены данные мониторинга уровня безработицы в Вологодской области (в рамках ежемесячного мониторинга населения РФ по вопросам экономической активности, занятости и безработицы). Очевидно, что уровень безработицы не может за один месяц снизиться с 11,8 % до 6 % (как в августе – сентябре 2009 г.) или подняться с 7,7 % до 10,8 % (как в январе – феврале 2010 г.) и тут же упасть до 7 % (как в феврале – марте 2010 г.). Таким образом, приведенные данные свидетельствуют о статистической недостоверности полученных оценок, т.е. эти выборки не являются репрезентативными. Рис. 1 наглядно это демонстрирует.

Как уже говорилось ранее, простейшим методом сглаживания является метод скользящего среднего [1]. Он заключается в том, что данные выборочных обследований за несколько последовательных месяцев (в простейшем случае – за три месяца) объединяются в одну выборку, по этой укрупнённой выборке рассчитывается среднее значение показателя, и оно условно относится к среднему месяцу. Для большинства задач социально-экономического мониторинга выборка, построенная путём объединения выборок трёх последовательных месяцев, достаточно представительна, и построенный этим методом временной ряд оказывается достаточно гладким. Однако метод скользящего среднего имеет существенный недостаток – чтобы рассчитать значение скользящего среднего за текущий месяц, необходимы данные выборочного обследования за следующий месяц. В работе предлагается метод сглаживания, который свободен от этого недостатка.

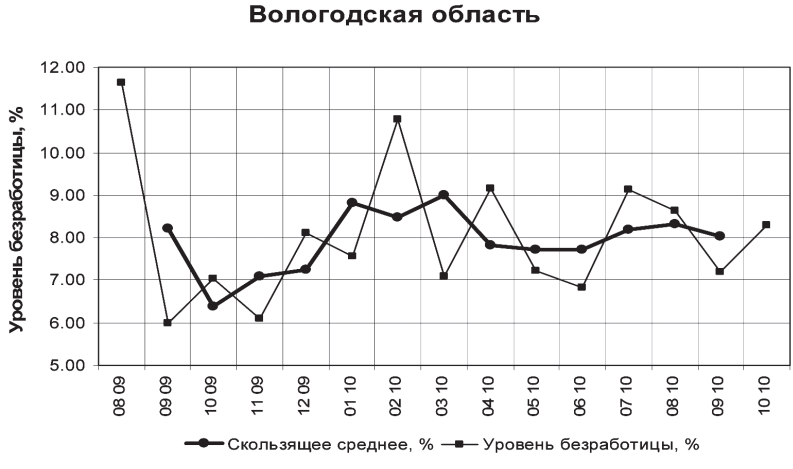


Рис. 1. Уровень безработицы в Вологодской области

2. МЕТОД СТРУКТУРНОЙ ГРУППИРОВКИ ОБЪЕКТОВ

Основная идея предлагаемого ***метода структурной группировки объектов (МСГО)*** состоит в том, что для повышения надёжности оценки исследуемого показателя производится усреднение не по времени, а по ансамблю объектов мониторинга. Это является ключевым отличием предлагаемого метода от других методов сглаживания временных рядов (например, от метода скользящего среднего, когда в одну группу объединяются выборки наблюдений за разные месяцы для одного и того же объекта). При использовании же метода МСГО в одну группу для дальнейшего расчёта оценки искомого показателя анализируемого i-го объекта объединяются выборки, полученные в одном и том же месяце для нескольких объектов, близких в определённом смысле по динамике исследуемого показателя к i-му объекту.

Ниже МСГО описан как метод оценки исследуемого показателя y для i-го объекта в k-м месяце текущего года. Далее этот объект называется i-эталонным, а k-й месяц – расчётным**.** Метод включает 3 этапа:

Этап 1. Производится сглаживание помесячных данных мониторинга i-эталонного объекта, для чего используется процедура трёхточечного скользящего среднего.

Этап 2. На этом этапе используется разработанный авторами статьи алгоритм i-эталонной классификации [5] (подробное описание см. раздел 2.2). С помощью этого алгоритма формируется класс объектов, близких (в определённом смысле) к i-эталонному объекту по динамике исследуемого показателя. Выборки вошедших в этот класс объектов для каждого момента времени из рассматриваемого диапазона мониторинга объединяются, то есть эти объекты рассматриваются как один виртуальный объект, ассоциируемый с i-эталонным объектом.

Этап 3. На базе объединённых выборок для виртуального объекта с помощью процедуры масштабирования находится искомая оценка исследуемого показателя для i-эталонного региона по состоянию на расчётный месяц.

2. 1. ФОРМИРОВАНИЕ ВИРТУАЛЬНОГО ОБЪЕКТА ДЛЯ *I*-ЭТАЛОННОГО ОБЪЕКТА

Для формирования виртуального объекта используются выборочные данные за каждый из 13 месяцев (расчётный месяц и за 12 месяцев, предшествующих расчётному). Выбор временного ряда такой длины диктуется следующим. Для того чтобы определить, являются ли два объекта близкими по динамике исследуемого показателя y, необходимо сопоставить его значе-

ния в двух объектах за период не меньше года (так как сезонные изменения y могут проявляться в течение всего года). Поскольку при формировании виртуального объекта используются не только исходные данные, но и их сглаженные значения (полученные с помощью процедуры скользящего среднего), то необходимы данные за дополнительный месяц в начале временного ряда. В качестве оценки скользящего среднего для расчётного месяца берется полусумма значений показателя за расчётный и предыдущий месяцы.

Формирование виртуального объекта производится с помощью предлагаемого в работе итерационного алгоритма i-эталонной классификации динамических объектов (траекторий) [5].

2.2. АЛГОРИТМ *I*-ЭТАЛОННОЙ КЛАССИФИКАЦИИ ДИНАМИЧЕСКИХ ОБЪЕКТОВ

Дадим вначале формальную постановку задачи i-эталонной классификации на примере задачи помесячного мониторинга исследуемого показателя y для N объектов. Пусть в процессе мониторинга для каждого l-го объекта получены n_l выборочных значений показателя y как за расчётный месяц, так и за каждый из 12 месяцев, предшествующих расчётному. Задача i-эталонной классификации состоит в разбиении по этим данным N объектов на такие 2 класса (i-эталонный и фоновый классы), чтобы выбранный критерий качества классификации $J_{эт}$ принимал максимальное значение. В работе в качестве критерия $J_{эт}$ используется значение коэффициента корреляции r_j между двумя временными рядами (векторами) – рядом оценок показателя y, полученных по объединённой выборке для объектов, отнесённых к i-эталонному классу (виртуальному объекту) $y_{вирт,k}$, как за расчётный месяц, так и за каждый из 11 месяцев, предшествующих расчётному; и рядом скользящих средних помесячных данных за тот же период времени мониторинга i-эталонного объекта $y_{cc}^{этал}$, то есть величина

$$J_{эт} = r_k = r(y_{вирт,\,k}, y_{cc}^{этал}). \qquad (1)$$

В формуле (1) величина k – это номер набора регионов, составляющих i-эталонный класс, однозначно определяющий номера регионов, входящих в этот набор. Тогда к i-эталонному классу относится такой набор объектов под номером m, который доставляет максимум критерию (1), то есть $m = \arg\max_k J_{эт}(k) = \arg\max_k r(y_{вирт,\,k}, y_{cc}^{этал})$. Все остальные

объекты относятся к фоновому классу. Отметим, что критерий качества классификации $J_{эт}$ (1) отличается от всех остальных тем, что в явном виде не зависит от объектов, отнесённых к фоновому классу. Очевидно, что для получения глобально оптимальной в смысле (1) i-эталонной классифика-

ции необходимо произвести полный перебор всех возможных поднаборов объектов из исходного набора N объектов.

В работе предложен эвристический алгоритм максимизации (1), с точки зрения основной идеи похожий на алгоритм пошаговой регрессии. Он представляет собой итерационную процедуру, на каждом шаге которой к i-эталонному классу при определённых условиях присоединяется наиболее близкий к нему объект из тех, которые к этому шагу не вошли в i-эталонный класс.

Для удобства описания алгоритма объектам присваиваются номера в том порядке, в котором они относятся к i-эталонному классу: i-эталонному объекту присваивается номер 1, следующему объекту, отнесённому к i-эталонному классу, – номер 2 и т.д.

Рассмотрим (j+1)-й шаг алгоритма. К началу (j+1)-го шага i-эталонный класс (виртуальный объект) включает j объектов, отнесённых к нему на предыдущих шагах, и представлен следующей информацией:

1. Временной ряд значений скользящего среднего оценок показателя y для i-эталонного объекта за 12 месяцев, предшествующих расчётному месяцу, а также оценка скользящего среднего для расчётного месяца, равная полусумме исходных значений оценок показателя расчётного и предыдущего месяца для этого объекта (далее этот временной ряд будем называть i-эталоном):

$$y^{i}_{эт ал} = y^{i}_{cc} = \left(y^{i\,(2)}_{cc},\ldots,y^{i\,(12)}_{cc},\hat{y}^{i\,(13)}_{cc}\right). \qquad (2)$$

2. Временной ряд (вектор) значений оценок показателя y, полученных по объединённой выборке для объектов, отнесённых к (j+1)-му шагу алгоритма к i-эталонному классу (виртуальному объекту) $y^{j}_{вирт}$, как за расчётный месяц, так и за каждый из 11 месяцев, предшествующих расчётному:

$$y^{j}_{вирт} = (y^{j(2)}_{вирт},\ldots y^{j(13)}_{вирт}). \qquad (3)$$

3. Коэффициент корреляции между временными рядами (2) и (3):

$$r_j = r(y^{j}_{вирт}, y^{i}_{эт ал}). \qquad (4)$$

На (j+1)-м шаге алгоритма из всех объектов, ещё не вошедших в i-эталонный класс (виртуальный объект), выбирается такой s-й объект, добавление которого к i-эталонному классу доставляет максимум коэффициенту корреляции r_{j+1}:

$$r_{j+1}(s) = \max_{l} x\left(r_{j+1}(l)\right). \qquad (5)$$

Если $r_{j+1} \geq r_j$ (коэффициент корреляции после включения s-го объекта в i-эталонный класс (виртуальный объект) не уменьшился), то этот объект добавляется к i-эталонному классу, ему присваивается номер (j+1) и алгоритм переходит к следующему шагу.

Если же $r_{j+1} < r_j$ (коэффициент корреляции уменьшился), то работа алгоритма заканчивается.

На рис. 2 в качестве примера приведена иллюстрация работы алгоритма *i*-эталонной классификации динамических объектов на примере Вологодской области. Как видно из рисунка, в виртуальный регион, ассоциированный с Вологодской областью, вошло 13 регионов (включая саму Вологодскую область). В процессе добавления новых регионов в группу коэффициент корреляции возрастал в пределах от 0,2 до 0,97. На рис. 2 объекты расположены по оси абсцисс в порядке их включения в виртуальный регион.

2.3. ПРОЦЕДУРА МАСШТАБИРОВАНИЯ

Несмотря на то, что временные ряды (2) и (3) по форме могут почти не отличаться друг от друга (в прикладных задачах коэффициент корреляции между соответствующими временными рядами, как правило, больше 0,9), их средние значения и масштаб могут заметно отличаться. Такое смещение и изменение масштаба объясняется тем, что в качестве меры близости временных рядов при формировании виртуального объекта используется значение коэффициента корреляции. А смещение на константу и изменение масштаба не меняют этого значения.

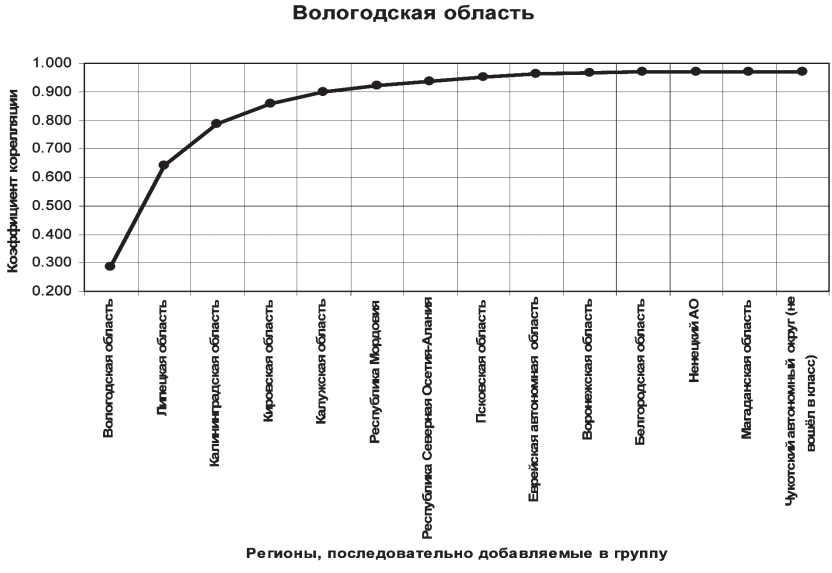


Рис. 2. Регионы, вошедшие в виртуальную группу для Вологодской области

Для демонстрации такого смещения на рис. 3 показан временной ряд уровня безработицы в виртуальном регионе, сформированном для Вологодской области, и для сравнения – временной ряд скользящего среднего этого показателя для Вологодской области. На этом примере (который является типичным) видно, что кривая уровня безработицы в виртуальном регионе достаточно гладкая. Как правило, она оказывается даже более гладкой, чем скользящее среднее для исходного временного ряда расчётного региона, потому что объём выборки по виртуальному региону обычно больше, чем объём выборки по расчётному региону за три месяца.

Таким образом, к *i*-эталонному классу могут относиться объекты, близкие к расчётному по характеру сезонных изменений значений оценок показателя *y*, но заметно отличающиеся по абсолютной величине этих оценок и с большей или меньшей амплитудой их колебаний (масштаба этой величины). Для того чтобы устранить полученное в результате этого смещение и изменение масштаба, производится линейное преобразование временного ряда (3), которое далее будет называться процедурой масштабирования.

Цель процедуры масштабирования – с помощью линейного преобразования временного ряда (3) (т.е. смещением на константу и изменением масштаба) так «совместить» его с временным рядом (2), чтобы сумма квадратов разностей между этими рядами по всем месяцам была минимальной.

Формально эта задача формулируется следующим образом: требуется найти такие константы b_0 и b_1 линейной регрессии $y^{i}_{эталл}$ на $y^{j}_{вирт}$, чтобы

$$\Delta y = \sum_{l=2}^{13} \left[y_{cc}^{i\,(l)} - \left(b_1 y_{вирт}^{j(l)} + b_0 \right) \right]^2 \rightarrow \min, \qquad (6)$$

где величины $y_{cc}^{i\,(l)}$ и $y_{вирт}^{j(l)}$ определены в (2) и (3) соответственно. Задача

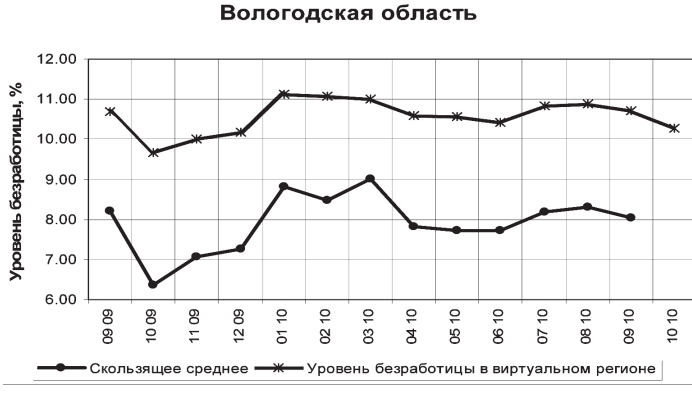


Рис. 3. Уровень безработицы в виртуальном регионе и скользящее среднее уровня безработицы в Вологодской области

нахождения оптимальных коэффициентов линейной регрессии (6) решается с помощью стандартной процедуры метода наименьших квадратов.

Результат решения этой задачи для рассматриваемого примера Вологодской области представлен на рис. 4.

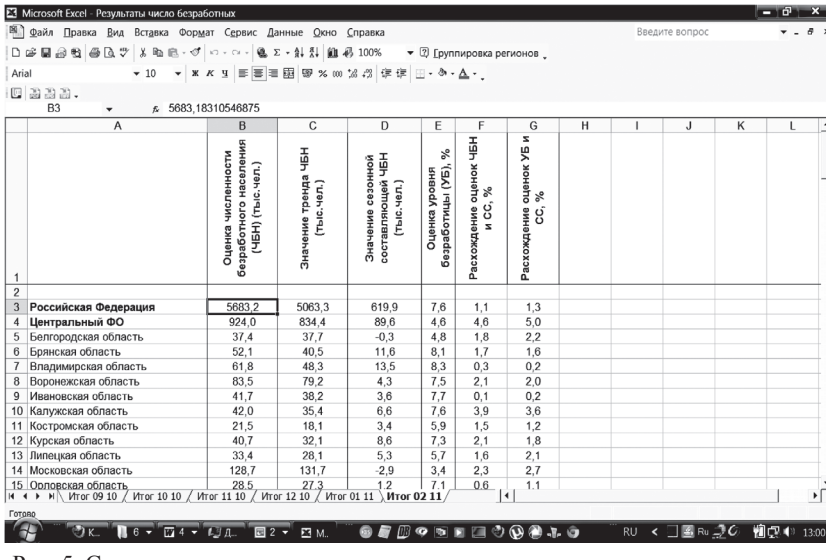
Как видно из представленного рисунка, в результате применения процедуры масштабирования временной ряд уровня безработицы в виртуальном регионе для Вологодской области и временной ряд скользящего среднего этого показателя оказались практически совмещены. Это показывает высокую точность разработанного метода.

3. РЕЗУЛЬТАТЫ ПРИМЕНЕНИЯ МЕТОДА СТРУКТУРНОЙ ГРУППИРОВКИ ОБЪЕКТОВ В ПРАКТИЧЕСКИХ ЗАДАЧАХ

В рамках применения на практике предлагаемого в статье метода МСГО была рассмотрена задача корректировки (сглаживания) оценок показателей экономической активности населения по субъектам РФ в условиях малых выборок. В настоящее время по вопросам экономической активности, занятости и безработицы ежемесячно опрашивается около 69 тыс. человек в возрасте 15–72 года (около 33 тыс. домашних хозяйств), или 0,06 % населения данного возраста. Объём месячной выборки обеспечивает представительные данные только в целом по РФ и некоторым крупным (по численности населения) субъектам РФ. Для двух третей субъектов РФ объемы месячной выборки являются недостаточными для получения представительных данных по показателям безработицы. Однако для эффективного мониторинга

Рис. 4. Результат масштабирования

требуются оценки для каждого конкретного месяца по каждому субъекту РФ, при этом в оценках должны быть учтены колебания, вызванные фактором сезонности, эффектом размещения выборки. Разработанный метод МСГО был успешно использован в Федеральной службе государственной статистики (Росстате) при решении задачи разработки системы алгоритмов и программных средств для автоматизации процедуры достоверного оценивания показателей мониторинга экономической активности населения, занятости и безработицы по субъектам РФ, формируемых по итогам месячных обследований населения по проблемам занятости [6]. В результате применения этого метода для каждого из 83 регионов РФ в период с сентября 2010 г. по октябрь 2011 г. были проведены расчёты ежемесячных оценок значений следующих показателей: численности экономически активного населения, уровня экономической активности, численности безработного населения, уровня безработного населения, численности занятого населения, уровня занятости населения. На базе предлагаемого в статье метода МСГО была разработана программа автоматического расчета оценок искомого показателя мониторинга как встраиваемый модуль для работы в программе Microsoft Excel. На рис. 5 представлена таблица, получаемая при расчете оценок численности безработного населения для всех регионов РФ за расчетный месяц. В программе также автоматически рассчитываются значения сезонной составляющей и тренда оценки показателя, как и сравнение оценок, полученных методами скользящего среднего и МСГО. При необходимости программа выводит на экран графики изменения динамики



	A	B	C	D	E	F	G	H	I	J	K	L
1		Оценка численности безработного населения (ЧБН) (тыс.чел.)	Значение тренда ЧБН (тыс.чел.)	Значение сезонной составляющей ЧБН (тыс.чел.)	Оценка уровня безработицы (УБ), %	Расхождение оценок ЧБН и СС, %	Расхождение оценок УБ и СС, %					
2												
3	**Российская Федерация**	5683,2	5063,3	619,9	7,6	1,1	1,3					
4	**Центральный ФО**	924,0	834,4	89,6	4,6	4,6	5,0					
5	Белгородская область	37,4	37,7	-0,3	4,8	1,8	2,2					
6	Брянская область	52,1	40,5	11,6	8,1	1,7	1,6					
7	Владимирская область	61,8	48,3	13,5	8,3	0,3	0,2					
8	Воронежская область	83,5	79,2	4,3	7,5	2,1	2,0					
9	Ивановская область	41,7	38,2	3,6	7,7	0,1	0,2					
10	Калужская область	42,0	35,4	6,6	7,6	3,9	3,6					
11	Костромская область	21,5	18,1	3,4	5,9	1,5	1,2					
12	Курская область	40,7	32,1	8,6	7,3	2,1	1,8					
13	Липецкая область	33,4	28,1	5,3	5,7	1,6	2,1					
14	Московская область	128,7	131,7	-2,9	3,4	2,3	2,7					
15	Орловская область	28,5	27,3	1,2	7,1	0,6	1,1					

Рис. 5. Скриншот окна программы для расчета оценок показателя мониторинга

оценок расчётного показателя по месяцам как для каждого отдельного региона (объекта), так и для РФ в целом.

Ввиду того, что в Росстате до введения в эксплуатацию метода МСГО для расчёта оценок показателей использовался метод скользящего среднего (МСС), для проверки эффективности МСГО было проведено сравнение оценок, полученных этими двумя методами. Результаты этих расчётов позволяют сделать следующие выводы. Оценки соответствующих показателей, полученные МСС и МСГО, очень близки (около 2 % от величины оцениваемого параметра для самых «проблемных» регионов). Ошибки МСС – это ошибки интерполяции. Ошибки МСГО связаны с неоднородностью выборки. При разных источниках ошибок результаты получаются достаточно близкими, это говорит об эффективности использования МСГО для достоверной оценки уровня соответствующего параметра. Однако МСГО имеет решающее преимущество: позволяет получать оценки уровня анализируемого параметра сразу после получения данных выборочного обследования.

ЛИТЕРАТУРА

1. *Бокс Дж., Дженкинс Г.* Анализ временных рядов // Прогноз и управление. Вып. 1, 2. – М.: Мир, 1974.

2. *Judge G. G., Griffits W. E., Hill R. C., Lutkepohl H., Lee Tsoung-Chao.* The Theory and Practice of Econometrics. Second edition. NY: John Willey and Sons, 1985.

3. *Introduction* to Seasonal Adjustment, DEMETRA+. − URL: http://circa.europa.eu/irc/dsis/eurosam/info/data/ demetra.htm

4. *Бауман Е.В., Дорофеюк А.А., Дорофеюк Ю.А.* Методы динамического структурного анализа многомерных объектов // Сборник трудов 4-й международной конференции по проблемам управления (МКПУ-IV). − М.: ИПУ РАН, 2009. – С. 338−343.

5. *Дорофеюк Ю.А., Дорофеюк А.А., Лайкам К.Э., Чернявский А.Л.* Алгоритмы эталонной кластеризации в задаче повышения достоверности статистических показателей в условиях нерепрезентативных выборок // Управление развитием крупномасштабных систем (MLSD'2011): Труды Пятой международной конференции. − М.: ИПУ РАН, 2011. − Т. I. – С. 268−275.

6. *Лайкам К.Э., Дорофеюк А.А., Дорофеюк Ю.А., Чернявский А.Л.* Классификационные методы коррекции результатов мониторинга социально-экономических показателей в условиях нерепрезентативных выборок // Вопросы статистики. − 2011. − №5. – С. 13−18.