

**ПРОБЛЕМА МОДЕЛИРОВАНИЯ МАССОВЫХ СОБЫТИЙ В СОЦИУМЕ НА
ОСНОВЕ ФРАГМЕНТАРНЫХ ДАННЫХ****Д. Н. Верзилин, Т. Максимова, А. Титов (Санкт-Петербург)****Состояние социума, массовые события и проблема фрагментарности данных**

Необходимым условием эффективного управления социально-экономическим развитием страны является наличие полной и достоверной информации о текущем состоянии общества и событиях, определяющих изменения этого состояния. Мониторинг массовых событий позволяет оценить текущее состояние социума, определить закономерности его динамики, выявить условия, при которых возможна потеря равновесного состояния. Оценивание состояния социума осуществляется по результатам мониторинга массовых событий. Массовые события изменяют состояния элементов социума. Элементы социума представляют собой как отдельных индивидуумов, так и их совокупности, созданные на принципах организации и самоорганизации. Движение элементов социума – это процесс изменения их состояния. Социум находится в динамическом равновесии при постоянном движении. Примерами массовых событий для индивидуумов являются рождение, поступление в учебное заведение, его окончание, поступление на работу, увольнение с работы, начало заболевания, выздоровление, утрата трудоспособности, смерть, для организаций – создание (регистрация), банкротство, ликвидация, слияние предприятий и т.д. Элементы социума обладают характеристиками, оказывающими влияние на возможность и частоту наступления массовых событий. Примеры таких характеристик – территориальная принадлежность элемента, половозрастная группа, профессиональный и квалификационный статус индивидуума, форма собственности, объема активов предприятия и т.п. Массовые события и состояния элементов социума регистрируются государственными и негосударственными организациями (лечебно-профилактическими учреждениями, учреждениями образования, бюро медико-социальной экспертизы и др.), а также в ходе специальных мероприятий, таких как социологические опросы, переписи и т.п. Сведения о массовых событиях обобщаются государственными органами различных уровней управления и ведомственной принадлежности (федеральными органами государственной власти, органами государственной власти субъектов Российской Федерации, органами местного самоуправления, органами управления здравоохранением, социальным обеспечением и др.).

Анализ состояния социума сопряжен с решением проблемы фрагментарности и необходимости реконструкции данных. Сведения о массовых событиях неизбежно имеют фрагментарный характер, что вызвано следующими основными причинами.

1. Процедуры регистрации массовых событий, применяемые в различных ведомствах РФ и в различных странах, не согласованы по временным интервалам подсчета массовых событий, по учитываемым характеристикам элементов социума (например, используются различные группировки по возрастным и социальным группам), по учитываемым характеристикам самих событий.

2. Процедуры регистрации массовых событий претерпевают изменения с течением времени.

3. Информация о массовых событиях утрачивается при сборе и обобщении данных на различных уровнях управления. В частности, утрачивается детализация сведений об элементах социума и характеристиках событий. Применяются смеси и свертки вероятностных распределений количественных характеристик элементов социума, вычисляются итоговые значения числа событий по их категориям и интервалам времени с утратой значений исходных слагаемых.

Таким образом, весьма актуальными являются задачи математического обобщения, строгой формулировки, формальной постановки проблемы реконструкции социально-экономических данных, разработки инструментария их реконструкции на основе математико-статистического и имитационного моделирования массовых событиях в социуме и процессов их регистрации.

Состояние математико-статистической методологии анализа социально-экономических данных

Развитие в Российской Федерации и за рубежом математико-статистической методологии в интересах анализа социально-экономических данных сосредоточено на следующих основных проблемах.

Разработка новых подходов к построению устойчивых статистических процедур, т.е. процедур, которые мало зависят от незначительных искажений анализируемых данных и принятых модельных допущений об их природе. Исследованиям в этой области среди отечественных ученых занимались Л.Д.Мешалкин, Б.Т.Поляк, Б.Т.Титаренко, А.М.Шурыгин, Ю.С.Харин, А.Б.Цыбаков [18, 10, 14, 20, 21, 19].

Развитие теории и методологии статистических методов классификации многомерных наблюдений. Наиболее значимы в этой области труды С.А.Айвазяна, И.С.Енюкова, Л.Д.Мешалкина, В.М.Бухштабера [9]. Предлагаемый С.А.Айвазяном и В.М.Бухштабером подход к построению общей теории автоматической классификации предполагает представление всего множества алгоритмов автоматической классификации в виде иерархической структуры. На верхнем уровне находится универсальная математическая модель, компоненты которой образуют средство для единообразной постановки задач автоматической классификации, описания алгоритмов их решения и исследования их свойств. Переход на более низкие уровни происходит за счет конкретизаций, наполняющих компоненты структуры модели информацией о характере данных, конечной цели классификации, априорных гипотезах, результатах предварительной обработки и т.п.

Статистические выводы в условиях растущей размерности наблюдений. В 60-х годах прошлого столетия эта проблема была озвучена А.Н.Колмогоровым. В настоящее время концепция "добычи данных" (Data mining) приобретает все большее распространение как инструмент для управления деловой информацией в тех случаях, когда предполагается, что из имеющихся данных можно будет извлечь знания для принятия решений в условиях неопределенности. Термин "добыча данных" объединяет совокупность методов аналитической обработки больших массивов данных с целью выявить в них значимые закономерности изменения переменных и связи между переменными. Процедура добычи данных включает три основных этапа: 1) подготовка данных, в т.ч. очистка, преобразование данных и т.п.; 2) построение, анализ различных моделей и выбор наилучшей; 3) применение модели к новым данным для осуществления прогноза. Указанное направление получило широкое развитие в трудах зарубежных ученых [1, 2, 3, 4, 5, 7, 8]. В этой связи разработчики экономико-математических моделей все чаще ставят вопрос об актуальности проблем построения сложных динамических моделей, включающих в себя хозяйствующих субъектов макроуровня и агентов микроуровня, поведение которых ограничено рационально. Эти проблемы можно решить путем построения агентно-ориентированных моделей, представляющих собой новое научное направление.

Развитие концепций и методологии имитационного моделирования и разработка компьютерных моделей общества, которые объединяют разнообразные по идейному подходу, по уровню детальности, по целям исследования модели. В частности, с использованием принципов агентно-ориентированного (agent-based) моделирования раз-

рабатываются компьютерные модели, которые включают в себя достаточно большое число агентов с определенным набором свойств для проведения симуляций различных событий и явлений. Содержательно агент – это человек. В агрегированных версиях это может быть совокупность однородных по какому-то признаку людей. Используется также известный прием интерпретации агента как представителя (средний человек). Конечная цель создания таких моделей – отследить влияние флуктуации агентов, действующих на микроуровне, на показатели макроуровня. В качестве моделируемых событий рассматриваются различные аспекты деятельности людей (поиск работы, распределение бюджета и т.д.). По данному предмету выходит множество зарубежной и отечественной литературы [11, 16, 20]. Наиболее обширным журналом, находящимся в свободном доступе, является электронный интернет-журнал, посвященный прикладным проблемам имитационного моделирования в общественных науках JASSS [6].

Разработка новых принципов и методов построения эконометрических зависимостей, в частности принципа максимальной согласованности [15], который позволяет полнее учесть разнородную исходную информацию о виде зависимости, результатах наблюдений и характере их ошибок и дает возможность решать задачи построения зависимости между характеристиками экономических объектов, используя теоретическую информацию и данные об их наблюдаемых значениях.

Анализ результатов научных исследований в области количественной обработки данных о массовых событиях, разработки моделей и сценариев функционирования и развития социума свидетельствует о достаточно малом объеме работ, отражающих проблемы построения моделей и сопоставления результатов моделирования в условиях фрагментарности данных. Частные проблемы имитационного моделирования массовых событий, согласования статистических данных и совершенствования форм статистической отчетности рассматриваются в ряде работ, посвященных вопросам анализа состояния социально-экономических систем и выработки управленческих решений [22, 23, 24, 25, 26, 27, 28, 29, 30, 31].

Проблемы и виды реконструкции данных о массовых событиях.

Предлагаются следующие виды реконструкции данных.

По возможности сбора дополнительной информации следует различать: 1) реконструкцию данных с целевыми запросами сведений у организаций – регистраторов массовых событий; 2) реконструкцию данных с проведением специальных исследований (социологических опросов, обработки первичных документов); 3) реконструкция данных без получения дополнительных сведений.

По видам реконструированных данных следует различать: 1) оценку значений показателей, характеризующих массовые события, произошедшие на временных интервалах с пропущенными наблюдениями (реконструкция динамики показателей); 2) восстановление отсутствующих показателей (реконструкция взаимной зависимости показателей); 3) оценку характеристик элементов социума (реконструкция популяций); 4) оценку характеристик массовых событий (реконструкция свойств событий).

Источники привлекаемых сведений при реконструкции данных: 1) результаты целевых запросов и специальных исследований; 2) данные, полученные для смежных временных интервалов; 3) данные, полученные для выборки из целевой совокупности (подпопуляции); 4) данные, полученные для выборки, внешней по отношению к целевой совокупности.

Общие подходы, составляющие основу процедур реконструкции данных (перечислены в порядке возрастания неопределенности исходных данных, сложности моделей и снижения точности прогноза состояний социума): 1) использование инвариантов совокупностей событий и точных функциональных зависимостей между показателями, характеризующими массовые события; 2) использование статистически значимых за-

зависимостей между показателями; 3) описание социально-экономических систем в виде систем потоков (применение моделей системной динамики); 4) описание социально-экономических систем в виде систем массового обслуживания, использование математических моделей, непосредственно описывающих наступления массовых событий, с учетом логико-временных зависимостей между ними (применение имитационного моделирования).

Применение последних двух подходов в большей степени направлено на построение возможных сценариев изменения состояния социума, определение ситуаций, в которых возможна утрата равновесного состояния, и в меньшей степени позволяет оценить количественные характеристики состояния.

Необходимо решить ряд теоретических проблем реконструкции данных и частных прикладных задач.

Теоретические проблемы: 1) разработка новых подходов к построению процедур реконструкции данных, робастных к незначительным искажениям данных и модельным допущениям; 2) оценка точности реконструированных данных при многоэтапной реконструкции; 3) исследование положительных и отрицательных причинно-следственных связей между процессами наступления массовых событий, в частности, построение моделей реакции социума на целенаправленные воздействия (перераспределение ресурсов); 4) построение моделей согласованного и самоорганизующегося поведения элементов социума для ситуаций, в которых изменение состояния его элементов зависит от состояния других элементов (развитие эпидемий, поведение на потребительских рынках, функционирование социальных сетей).

Прикладные задачи: 1) определение наиболее информативных показателей, характеризующих состояние социума и основанных на свойствах массовых событий; 2) определение для показателей пороговых значений, при достижении которых возможна утрата социумом текущего динамического равновесия; 3) оптимизация процессов регистрации сведений о массовых событиях по критерию минимум затрат (временные и денежные ресурсы) – максимум информации (полнота и точность реконструированных данных) и, как следствие, межведомственная согласованность официальных форм статистической отчетности, оптимальный синтез целевых запросов.

Проведение исследований в следующих направлениях обеспечит решение поставленных задач

- Построение комплекса моделей для оценивания на основе фрагментарных данных распространенности заданных состояний элементов социума, интенсивности появления событий, связанных с переходом элементов социума в заданные состояния, продолжительности пребывания элементов в заданных состояниях.
- Разработка общего алгоритма оценивания перечисленных характеристик.
- Построение комплекса моделей для оценивания на основе фрагментарных данных реакции социума на целенаправленные воздействия.
- Разработка алгоритма определения наиболее информативных показателей, описывающих управляющие воздействия и реакцию социума.
- Разработка методики оценки точности реконструированных данных при многоэтапной реконструкции.
- Разработка процедур реконструкции данных, робастных к незначительным искажениям данных и модельным допущениям.
- Построение комплекса моделей согласованного и самоорганизующегося поведения элементов социума.
- Разработка алгоритма определения критических значений показателей состояния социума.

Выводы. Целесообразно разработать единый модельный комплекс, описывающий процессы наступления массовых событий в социуме, регистрации событий и обобщения информации о событиях. Создание такого модельного комплекса обеспечит повышение точности и надежности оценок состояния социума в результате мониторинга массовых событий.

Литература

1. **Berry M. J. A. & Linoff G. S.** Mastering data mining. New York: Wiley, 2000.
2. **Edelstein H. A.** Introduction to data mining and knowledge discovery (3rd ed). Potomac, MD: Two Crows Corp.
3. **Fayyad U. M., Piatetsky-Shapiro G., Smyth P. & Uthurusamy R.** (1996). Advances in knowledge discovery & data mining. Cambridge, MA: MIT Press, 1999.
4. **Han J., Kamber M.** Data mining: Concepts and Techniques. New York: Morgan-Kaufman, 2000.
5. **Hastie T., Tibshirani R., & Friedman J. H.** The elements of statistical learning: Data mining, inference, and prediction. New York: Springer, 2001.
6. Journal of artificial societies and social simulation (JASSS). <http://jasss.soc.surrey.ac.uk>.
7. **Weiss S. M. & Indurkha N.** Predictive data mining: A practical guide. New York: Morgan-Kaufman, 1997.
8. **Westpha, C., Blaxton T.** Data mining solutions. New York: Wiley, 1998.
9. **Айвазян С. А., Енюков И. С, Мешалкин Л. Д.** Прикладная статистика. М.: Финансы и статистика. 1983. Т.1: Основы моделирования и первичная обработка данных; 1985. Т. 2: Исследования зависимостей; 1989. Т. 3: (в соавт. с В.М.Бухштабером): Классификация и снижение размерности.
10. Алгоритмы многомерного статистического анализа и их применения. М.: Изд-во ЦЭМИ АН СССР, 1976.
11. **Бахтизин А. Р.** Агент-ориентированные модели экономики. М.: Экономика, 2008. 279 с.
12. **Болч Б., Хуань К. Дж.** Многомерные статистические методы для экономики. М.: Финансы и статистика, 1979.
13. **Жамбю М.** Иерархический кластер-анализ и соответствия. М.: Финансы и статистика, 1988.
14. Исследования по вероятностно-статистическому моделированию реальных систем. М.: Изд-во ЦЭМИ АН СССР, 1977.
15. **Клейнер Г. Б., Смоляк С. А.** Эконометрические зависимости: принципы и методы построения. М.: Наука, 2000. 104 с.
16. **Макаров В. Л., Бахтизин А. Р., Бахтизина Н. В.** CGE модель социально-экономической системы России со встроенными нейронными сетями. М.: ЦЭМИ РАН, 2005. 152 с.
17. Материалы Лаборатории искусственных обществ. <http://www.artsoc.ru/html/mission.htm>.
18. Многомерный статистический анализ в социально-экономических исследованиях: Ученые записки по статистике, Т. 26. М.: Наука, 1974.
19. Многомерный статистический анализ и вероятностное моделирование реальных процессов: Ученые записки по статистике. Т. 54. М.: Наука, 1990.
20. Прикладной многомерный статистический анализ: Ученые записки по статистике, Т. 33. М.: Наука, 1978.
21. Статистика. Вероятность. Экономика: Ученые записки по статистике. Т. 49. М.: Наука, 1985.
22. **Верзилин Д. Н., Черешнев В. В.** Статистический анализ риска в системах с критическими состояниями // Моделирование и анализ безопасности и риска в сложных

- системах: Труды международной научной школы МАБР-2004. СПб.: Изд-во ГОУ ВПО «СПбГУАП», 2004. С. 274–281.
23. **Верзилин Д. Н.** Методы анализа условий и качества функционирования социальных и медицинских систем в субъектах Российской Федерации. СПб.: Изд-во СПбГУЭФ, 2004. 175 с.
 24. **Уйба В. В., Рыжков Н. А., Верзилин Д. Н.** Процедуры формирования нормативных показателей целевой и финансово-хозяйственной деятельности медицинских учреждений федерального подчинения // Экономика здравоохранения. 2004. № 2. С. 43–46.
 25. **Верзилин Д. Н., Черешнев В. В.** Использование метода анализа соответствий при управлении производственно-сбытовой деятельностью предприятия // Экономика и математические методы. 2005. Т. 41. № 1. С. 56–64.
 26. **Верзилин Д. Н., Черешнев В. В.** Управление рисками в производственных и социальных системах на основе статистических данных // Моделирование и анализ безопасности и риска в сложных системах: Труды Международной научной школы МА БР- 2005. СПб., 2005. С. 384–389. (РАН).
 27. **Черешнев В. В., Верзилин Д. Н., Зайчик Е. С.** Имитационное моделирование конкурентного поведения производителя на потребительском рынке / Третья всероссийская научно-практическая конференция по имитационному моделированию и его применению в науке и промышленности «Имитационное моделирование. Теория и практика» ИММОД-2007. СПб., 2007.
 28. **Уйба В. В., Верзилин Д. Н.** Организация медицинской помощи при ликвидации последствий техногенных катастроф: аналитико-имитационное моделирование / Третья всероссийская научно-практическая конференция по имитационному моделированию и его применению в науке и промышленности «Имитационное моделирование. Теория и практика» ИММОД-2007. – СПб., 2007.
 29. **Верзилин Д. Н., Уйба В. В., Рыжков Н. А.** Анализ риска и контроль качества в задачах стратегического менеджмента: междисциплинарные связи / Международная научная школа “Моделирование и анализ безопасности, риска и качества сложных систем” МАБР-2007. СПб, 2007.
 30. **Верзилин Д. Н., Максимова Т. Г.** Эконометрика. Принятие решений на основе статистических данных (Учебное пособие). СПб.: Изд. Политех ун-та, 2008, 120 с.
 31. **Верзилин Д. Н., Черешнев В. В., Максимова Т. Г.** Управление сложными организационными системами: концепции, принципы, инструментарий. Екатеринбург: УрОРАН, 2009. 234 с.