Н.С. МАРТЫШЕНКО

Методическое обеспечение анализа поведения потребителей на региональном туристском рынке

Излагаются методические подходы и методика сбора и обработки первичных данных, характеризующих конъюнктуру рынка туристских услуг, а также рекомендации по применению методов многомерного статистического анализа для сегментирования потребительского рынка услуг туристского комплекса.

В последнее десятилетие существенно повысилась значимость Приморского края для экономики страны, в том числе такой перспективной отрасли как туризм. Многим жителям Дальнего Востока России и Сибири стали недоступны черноморские и прибалтийские курорты, рекреационные объекты Кавказа и Карпат в связи с существенно возросшими транспортными тарифами.

Одним из сдерживающих факторов развития туризма в Приморском крае является неразвитость туристской индустрии и туристской инфраструктуры, отсутствие навыков изучения рынка, методик исследования поведения потребителей и механизмов воздействия на них, слабое использование возможностей современных информационных технологий в решении этих задач. В настоящей работе рассматриваются некоторые пути преодоления проблем, существующих в развитии туристского комплекса региона.

Заинтересованность края в развитии туристской отрасли зависит от того, какой вклад она может внести в его развитие: поступления в бюджет региона, занятость и доходы населения, работающего в этой отрасли. Основной доход в бюджет приносят не те туристские фирмы, которые формируют туристский продукт и реализуют его потребителям, а предприятия социально-бытовой и туристской инфраструктуры региона, обслуживающие туристов (транспорт, гостиницы, рестораны и т.п.). Таким образом, задача анализа рынка туристских услуг усложнена нечеткостью границ сферы услуг, задействованной непосредственно в обслуживании туристов, тем более, что многие объекты можно расценивать как объекты двойного назначения.

Статистические данные, характеризующие сегодняшнее состояние рынка туристских услуг, свидетельствуют об устойчиво низкой активности потребителей. А при существующих потенциальных возможностях края по удовлетворению спроса на туристские услуги, вклад отрасли в экономику края нужно признать неадекватно малым (3% в ВРП). Этим

обусловлена цель анализа поведения потребителей туристских услуг: он должен обосновать действия и последовательность, в какой их необходимо предпринять, чтобы существенно повысить значимость туризма в экономике края.

В настоящее время в Приморье насчитывается более 250 туристских фирм и агентств. Коллективы фирм, как правило, очень малочисленны. Турфирмы со штатом сотрудников 3-5 человек составляют 37%, более 20 человек – всего 5% общего количества турфирм и агентств.

Около 63% туристских фирм принимают до 1 тыс. туристов в год, и они обслуживают всего 7% общего количества въездных туристов. С другой стороны, 7% туристских фирм, принимающих более 22 тыс. туристов в год, обслуживают около 42% общего числа всех туристов [1].

Структура предложений туристского продукта пока очень проста. Главным направлением поездок населения Приморского края является Китай – свыше 90% выездного потока, затем идут США, Корея и Япония. Понятно, что такой туризм в основном носит коммерческий характер: Китай – шоптуры, Япония – автомобильный бизнес. Наблюдается тенденция незначительного роста количества поездок жителей Приморья в Таиланд, США, Сингапур и Австралию.

Перспективы роста спроса на услуги туристской отрасли связаны с повышением уровня жизни населения, который согласно официальной статистике в последние годы имеет устойчивую положительную тенденцию.

Потребителей рынка туристских услуг можно разбить на три большие группы, или сегмента: иностранные туристы; туристы, прибывающие из других регионов страны; население края (потребление услуг во время отпусков и в свободное от работы время).

Анализ статистических данных показал, что общая численность туристов, прибывающих в край в течение года из других стран, в последние годы сохраняется примерно на одном уровне, а по некоторым странам (в том числе Китаю) даже снижается. Иностранные туристы наиболее выгодны для края, но резко изменить ситуацию в ближайшие годы, по всей видимости, не удастся, поскольку требования зарубежных туристов к уровню и качеству предоставляемых услуг намного выше, чем сегодня может предоставить туристская отрасль края.

Вторая группа потребителей – туристы, въезжающие из других регионов страны. Эта группа в настоящее время является наиболее перспективной, потому что очень многих жителей соседних регионов интересует отдых у моря, традиционно привлекательный для россиян.

Некоторые специалисты [3, 9], оценивая уровень развития туризма, акцентируют свое внимание на природно-климатических ресурсах, но мы связываем перспективу развития внутреннего и въездного туризма в Приморском крае с развитием материально-технической базы предприятий туристской индустрии и туристской инфраструктуры, повышением уровня обслуживания, а также с ведением разумной ценовой политики на товары и услуги, особенно туристского назначения.

В отличие от двух выше обозначенных групп въездных туристов масштаб третьей группы – жителей края – фиксирован. Хотя возможно-

сти внутреннего туризма ограничены численностью и доходами населения, он должен быть объектом первостепенного внимания, поскольку в дальнейшем может привести к росту и въездного туризма. Значимость этой идеи возрастает, если учесть, что в последние три года наметились явно положительные тенденции в росте благосостояния населения, а значит, и возможности совершать поездки по краю и потреблять товары и услуги регионального туристского комплекса.

Еще одним аргументом в пользу более глубокого исследования населения края как потребителя туристских услуг, является отработка методик анализа поведения потребителей в этой специфической отрасли. Особенно это важно для методик, основанных на применении сложных математических моделей и многомерного анализа данных. При всех различиях рассмотренных потребительских групп отлаженные методики могут быть в дальнейшем применимы и к въездному иностранному туризму. Сбор данных по этой категории потребителей намного сложнее и, соответственно, требует больших финансовых и временных затрат. Поэтому использование испытанных методик и технологий позволит более эффективно организовать сбор и обработку данных, необходимых для исследования въездного иностранного туризма.

Первичный материал о поведении потребителей формируется в процессе анкетных опросов потребителей туристских услуг. Обработку данных анкет можно разбить на три этапа: анализ достоверности и устранения грубых ошибок; предварительный анализ; многомерный анализ.

Прежде чем приступить к анализу данных, необходимо разработать систему сбора первичного материала. Исследователь всегда ограничен в источниках информации. Умение выявить источник и собрать сведения, которые возможно найти в условиях, сложившихся на момент проведения исследований, составляет основу практически любой научной работы. Наибольшую ценность представляют не те данные, которые удалось почерпнуть из внешних источников, отчетов и книг регистрации, а те, которые получены или изысканы самим исследователем путем целенаправленных действий. В этом случае исследователь самостоятельно определяет не только структуру и способ сбора информации, но и организует процесс сбора. При этом он понимает, каким образом эти данные будут представлены в компьютерной информационной системе, как будут использованы для проведения научного анализа. Все вместе, от определения информационной потребности и возможностей по сбору данных до системы представления данных на компьютере, можно определить как систему сбора первичных материалов.

Создание системы сбора информации, необходимой для решения задачи исследования поведения потребителей, – это сложный и трудоемкий процесс, предполагающий прохождение множества этапов на пути к поставленной цели. Разработку системы сбора первичных данных приходится проводить в условиях неполной информации и ограниченности ресурсов. Поиск решений осуществляется для слабо структурированных и не полностью формализованных задач. Эффективность работы такой системы будет тем выше, чем более полно в ней будут реализованы принципы системного подхода [5, 6].

Хотя сбор данных о поведении потребителей имеет единую цель (принцип конечной цели), в рамках одного опроса достигнуть ее невозможно. Каждый новый этап исследования (принцип поэтапности) дает новую информацию для организации работы на следующем этапе. Поэтому разрабатываемая система сбора должна отвечать принципу развития или открытости, который предполагает учет изменяемости системы и способности к развитию, расширению, накоплению информации. С расширением спектра решаемых задач и появлением новых функций возникает необходимость разработки новых модулей, совместимых с уже имеющимися.

Важнейшим принципом, которому должна отвечать как система сбора, так и система обработки информации, является принцип модульности. В процессе накопления данных уточняются формулировки вопросов, а сами вопросы объединяются в некоторые блоки или модули. Повторяемость модулей вопросов, используемых при анкетировании потребителей туристских услуг, позволяет выдержать основополагающее свойство системы – ее целостность [8].

Построение системы с учетом состояния внешней среды известно как удовлетворение принципа связности. Система сбора информации рассматривается как элемент (подсистема) более общей системы, имеющей своей целью решение главной задачи – исследование поведения потребителей.

После сбора анкетных данных необходимо провести анализ их достоверности. Этот этап имеет исключительно важное значение в практических исследованиях, потому что от степени достоверности данных зависит обоснованность выводов на всех последующих этапах. Для решения этой задачи используется такой метод многомерного статистического анализа, как рабастное оценивание.

Рабастные методы оценки данных используются для выявления и подавления грубых ошибок в первичных данных многомерных наблюдений. Основы теории рабастности были разработаны академиками А.Н. Колмогоровым, Н.В. Смирновым и Б.С. Ястремским. Свое дальнейшее развитие теория рабастности получила в работах зарубежных ученых А. Тьюки, Дж. Хьюбера.

Суть классических методов сводится к проверке сложных статистических гипотез на основе столь же сложных критериев (критерий Хьюберта, Граббса, Титьена-Мура), не нашедших широкого распространения в силу ряда причин, среди которых сложность их практического использования не является основной. Из анализа примеров использования таких методов можно сделать вывод, что они применяются в задачах с крайне ограниченными объемами выборок. Цена включения или отбрасывания одного объекта (наблюдения) достаточно высока, и поэтому для принятия решения недостаточно только интуиции исследователя и содержательного анализа. Решение должно быть подтверждено числом в вероятностном смысле. Соответственно, табулированные значения критических точек для специфических критериев рабастности, которые удалось обнаружить, в научной литературе приводятся для объемов выборки от 20 до 50. Программы расчета рабастных критериев не входят в состав

распространенных пакетов по обработке статистических данных [4]. Тем не менее проблема засорения первичных данных при обработке анкет существует и ее надо решать.

Для повышения достоверности анкетных данных можно найти более простые решения. Основная идея предлагаемого подхода состоит в построении некоторых фильтров, которые не дают точных результатов, а позволяют из большого количества наблюдений выделить «подозрительные» с точки зрения наличия грубой ошибки. Далее принятие решения отводится исследователю. Специфика исследований, основанных на анкетных данных, такова, что можно легко исключить 5–10 наблюдений из 2–3-х тысяч.

При массовом сборе анкетных данных исследователь сталкивается с необходимостью привлечения большого количества интервьюеров. Каждый интервьюер проводит опрос некоторой группы респондентов. Качество материала, собранного каждым интервьюером, может существенно различаться в силу разных причин, например недобросовестной работы некоторых интервьюеров, или их неумения контактировать с респондентом, или попытки интервьюера собрать данные в специфической контактной группе респондентов, негативно относящихся к опросу. Независимо от причины, исследователю часто выгоднее отказаться от всей серии анкет, существенно отличающихся от всех остальных, чем получить недостоверные данные. При нехватке статистического материала лучше организовать дополнительный опрос респондентов.

Проводить анализ выбросов по одному признаку (вопросу анкеты) при большом количестве анкет крайне неэффективно, особенно когда вопросов в анкете очень много. Практика показывает, что если анкета содержит недостоверные данные, то недостоверность не носит избирательного характера, а прослеживается по большинству вопросов. Поэтому лучше осуществлять фильтрацию по всем признакам. Однако поскольку ответы на различные вопросы могут представлять собой признаки, измеряемые в различных шкалах, такой анализ целесообразно делать по группе признаков, измеренных в одной шкале.

Предположим, что ответы на вопросы содержат m признаков, представленных в шкале отношений. Выборка является совокупностью k пакетов анкет, собранных различными интервьюерами. Одно наблюдение из r-го пакета (r – номер интервьюера) можно представить набором значений m признаков:

$$x_{i_r} = (x_{i_r 1}, x_{i_r 2}, ..., x_{i_r j}, ..., x_{i_r m}),$$

где $j = 1, 2, \ldots, m$ – номер признака; m – количество признаков; $r = 1, 2, \ldots, k$ – номер интервьюера, k – количество интервьюеров; $i_r = 1, 2, \ldots, n_r$ – номер анкеты в пакете одного интервьюера; n_r – объем пакета анкет интервьюера с номером r.

Тогда объем выборки, включающей все анкеты, будет:

$$n_0 = \sum_{r=1}^{k} n_r .$$

Задача состоит в том, чтобы из k пакетов анкет выделить тот, который имеет наибольшие отличия от остальных.

С этой целью последовательно для каждого пакета r (r = 1, 2, …, k) повторим следующую процедуру: рассчитаем средние значения m признаков по выборке за исключением пакета с номером r:

$$\overline{X}^{-r} = (\overline{x}_1^{-r}, \overline{x}_2^{-r}, ..., \overline{x}_j^{-r}, ..., \overline{x}_m^{-r}),$$

и средние значения признаков по пакету с номером r:

$$\overline{X}^{r} = (\overline{x}_1^{r}, \overline{x}_2^{r}, ..., \overline{x}_j^{r}, ..., \overline{x}_m^{r}).$$

Вычислим поэлементные модули разностей двух векторов средних:

$$\lambda_{rj} = \left| \overline{X}_j^{-r} - \overline{X}_j^{r} \right|.$$

Процедура, основанная на последовательном изъятии и восстановлении части выборки, называется скользящим экзаменом. Объединяем все отклонения λ_{rj} в одну матрицу λ размерности $k \times m$. На основании матрицы λ рассчитаем матрицу M той же размерности. Вычисления производятся по схеме: определяется максимум в каждом столбце матрицы λ, затем элементу матрицы M, соответствующему максимуму, присваивается значение единицы, всем остальным элементы матрицы M присваивается значение ноль. В результате построчного суммирования элементов матрицы M получим вектор оценок для каждого интервьюера:

$$\mu = (\mu_1, \mu_2, ..., \mu_r, ..., \mu_k).$$

Интервьюер с наибольшим значением μ_r будет иметь максимальный штраф, и поэтому его данные могут быть поставлены под сомнение. Теперь исследователь может сосредоточить свое внимание на отдельном пакете первичных данных, подвергнуть их дополнительному содержательному анализу, в результате которого он определит, является отклонение допустимым или нет. Это достаточно грубый фильтр. Он основан на предположении: если пакет анкет содержит недостоверную информацию, то большие отклонения от средних будут не только по одному признаку, но и по другим. Данную процедуру можно применить и в том случае, когда пакеты содержат по одному наблюдению. Тогда идентификатором интервьюера может выступить номер наблюдения в выборке.

Для признаков, измеренных в ранговых и номинальных шкалах, можно использовать фильтр, основанный на сравнении частотных рядов. К номинальным шкалам можно отнести и бинарные признаки. К этой шкале с помощью операции типологии могут быть преобразованы и ответы на открытые вопросы. Практика показывает их высокую информативность.

Предположим, что всего выделено m признаков. Тогда для расчета элемента вектор λ_r будем использовать формулу:

$$\lambda_{rj} = \sum_{i_r}^{f_r} \frac{(P^{-r}{}_{i_r j} - P^{r}{}_{i_r j})^2}{P^{-r}{}_{i_r j}},$$

где i_r = 1, 2,…, f_r – номер интервала частотного ряда r-го признака; f_r – количество интервалов в частотном ряду r-го признака; $P^{-r}{}_{i_r j}, P^{r}{}_{i_r j}$ – относительные частоты, рассчитанные по частотным рядам выборки без r-го пакета и только по выборке r-го пакета.

Поскольку количество разрядов в частотных рядах различных признаков может различаться, то в общей оценке интервьюера признаки должны браться с весовыми коэффициентами:

$$\mu_r = \sum_{j=1}^{m} \frac{f_j}{\sum_{j=1}^{m} f_j} \lambda_{rj}.$$

Необходимость весовых коэффициентов отпадает, если все признаки свести к бинарным шкалам. Соответственно, можно использовать второй фильтр на всех признаках, если предварительно перейти от шкалы отношений к ранговой или бинарной шкале.

Третий фильтр является наиболее точным, поскольку максимально использует информацию, содержащуюся в выборке. Ввиду громоздкого формального описания этого фильтра, изложим только основную идею. Фильтр основан на использовании идей распознавания образов. Вначале все признаки приводятся к бинарному представлению. Выборка считается разделенной на два класса. Первая выборка – анкеты одного пакета, вторая выборка – все остальные анкеты. Две выборки используются для построения эталонов классов (обучение). Затем проводится распознавание наблюдений обеих выборок, по результатам которого рассчитывается количество ошибок распознавания. Далее процедура повторяется по схеме скользящего экзамена. Пакет анкет, обеспечивающий наименьшую ошибку распознавания, можно считать более обособленным и, следовательно, он рассматривается как аномалия, требующая содержательного анализа. В этом случае могут быть изъяты не только наблюдения (анкеты), принадлежащие этому пакету, но и прочие наблюдения, отнесенные к аномальной выборке в процессе распознавания.

При содержательном анализе подозрительных данных, как правило, обращаются к первичному материалу (анкетам на бумажном носителе). Часто выбросы могут быть вызваны банальной ошибкой оператора или явной опиской респондента, автоматически перенесенной оператором в компьютер. Такие ошибки легко устранимы.

Рассмотренные выше фильтры предполагают, что все респонденты ответили на все вопросы анкет. На практике это выполняется крайне редко. Причин этому может быть несколько:

1) не все респонденты одинаково хорошо могут понимать вопрос анкеты. В этом случае респондент может просто отказаться отвечать на такой вопрос;

2) на некоторые вопросы респондент может отказаться отвечать, потому что считает информацию конфиденциальной;

3) респондент может просто по невнимательности пропустить вопрос;

4) респондент может и понимать вопрос, но просто не владеть требуемой информацией для ответа на него;

5) отсутствие данных в ответе на один вопрос может быть вполне компенсировано ответом на другой вопрос. Например, если нужно указать сумму денежных средств, затрачиваемую в отпускное время, отсутствие данных вполне может быть объяснимо, если респондент вообще не использовал отпуск;

6) респондент может не давать ответы на ряд вопросов при негативном отношении к опросу.

По крайней мере, пять причин отсутствия данных по одному признаку в наблюдении (ответ на один вопрос) не влияет на достоверность данных при ответах на другие вопросы. Все такие ситуации, хотя и создают дополнительные трудности исследователю, но вполне могут быть обработаны. Во-первых, в отдельных задачах анализа возможно использование не всех признаков и, следовательно, отсутствие данных в неиспользуемых признаках вообще не отражается на таких задачах. Во-вторых, некоторые позиции могут заполняться на основе содержательного анализа соответствующей анкеты. В-третьих, иногда отсутствие данных можно интерпретировать как значение, измеренное в порядковой шкале.

Шестая причина ставит под сомнение достоверность ответов и на все остальные вопросы. Такие анкеты должны быть по возможности выявлены с помощью специальных фильтров и удалены. Такой фильтр может быть построен на основе идей распознавания образов.

Для анализа многомерной выборки на предмет отсутствия данных может быть очень полезен следующий простой фильтр.

Матрица данных наблюдений (за исключением столбцов, содержащих все данные) преобразуется к бинарному виду (0 – есть данные, 1 – нет данных). Обозначим такую матрицу через B. Она имеет размерность $n \times m$, где n – количество наблюдений, а m – количество признаков. Произвольный элемент матрицы обозначим b_{ij}. Тогда каждой строке матрицы (ассоциированной с наблюдением) можно сопоставить некоторую величину:

$$\lambda_i = \sum_{j=1}^{m} \frac{1}{v_j} b_{ij},$$

где v_j – количество единиц в столбце j.

Весовой коэффициент $1/v_j$ вводится для ранжирования признаков. Если признак содержит множество незаполненных позиций (отсутствие данных), то отсутствие данных признака в отдельном наблюдении – событие не столь уж исключительное и имеет небольшой вес и наоборот. После расчета λ_i ($i = 1, 2, \ldots, n$) выборка может быть упорядочена по убыванию показателя λ_i. На первых позициях окажутся наблюдения, внушающие наибольшее беспокойство отсутствием данных. Анкеты, соответствующие таким наблюдениям, должны быть подвергнуты углубленному содержательному анализу. Если в процессе анализа возникает подозрение, что отсутствие данных вызвано шестой причиной, то такие анкеты должны быть исключены. Степень уверенности в необходимости исключения данных анкеты повышается, если в поле зрения исследователя попадает серия анкет, принадлежащих одному пакету.

Только после того, как исследователь убедился, что полученные данные заслуживают доверия и не содержат грубых ошибок, он может приступать к дальнейшему анализу анкетных данных.

На втором этапе проводится предварительный анализ данных, собранных в процессе анкетирования потребителей туристских услуг на региональном рынке. Как правило, производится построение частотных

рядов отдельных признаков, осуществляется формирование признаков на основании типологий, создаваемых по открытым вопросам анкет.

Среди задач исследования поведения потребителей ключевой является сегментирование потребителей, или по терминологии методов статистического анализа (МСА) задача многомерной классификации [2, 6]. Хотя этот раздел МСА в последние два десятилетия очень бурно развивался, находя все новые и новые области применения, задачи, основанные на использовании методов кластерного анализа, до сих пор больше являются прерогативой научных исследований, а не относятся к повседневной практике. Особенно это касается таких относительно новых объектов, как потребители туристских услуг. Однако с распространением специализированных пакетов программ по статистике появляется возможность приблизить сугубо научную задачу к практической работе.

В целях практического применения кластер-анализа для сегментирования потребителей экономисту недостаточно знания основ теории кластеризации и наличия соответствующих программных средств, необходима конкретная методика, которая прослеживала бы все этапы от зарождения идеи поиска сегментов потребителей до получения результата. В методике должны быть определены возможные проблемы, с которыми может столкнуться исследователь, решивший воспользоваться новым инструментом, и пути преодоления этих проблем.

На самом деле идея наличия классов однородных объектов возникает у исследователя еще до сбора первичных данных и разработки содержания вопросов анкет. Она основана на объективно существующем и наблюдаемом явлении в поведении потребителей, состоящем в предпочтении разными группами потребителей тех или иных услуг. Что объединяет одних, что разъединяет других, каковы закономерности формирования однородных групп потребителей? Самые общие идеи о причинах формирования групп потребителей должны зародиться у исследователя еще до сбора данных. При этом у него может быть несколько гипотез относительно формирования сегментов потребителей. Наличие нескольких гипотез даже желательно, поскольку, если исходная гипотеза одна, и в конце концов выяснится, что она не подтвердилась, все усилия по поиску сегментов окажутся напрасными, что может обнаружиться на последних этапах предпринятого исследования. Если говорить о потребителях туристских услуг, то задача выделения однородных сегментов еще более усложняется, поскольку туристская услуга комплексная и не имеет четко очерченных границ. Формулирование гипотез о формировании сегментов можно признать **первым этапом** на пути решения задачи сегментирования. Чаще всего на первых этапах исследователь не задается вопросом о количестве сегментов.

На втором этапе исследователь определяет для себя, какими признаками он может охарактеризовать возможные сегменты, пока не задумываясь, как он сможет получить информацию о них. Это желаемые характеристики потребителей.

На третьем этапе исследователь должен определить поле признаков или характеристик, которые он может реально получить в процессе анкетного опроса. На этом этапе он моделирует восприятие респондентами возможных вопросов. При этом он должен учитывать доступные

ему средства организации опросов. Как правило, эти характеристики очень трудно получить, поэтому исследователь должен сформулировать целый ряд вопросов, которые, по его мнению, в комплексе отражают уровни желаемой характеристики. На этом же этапе исследователь определяет метод сбора информации.

На четвертом этапе происходит формирование содержания анкеты. Здесь опять исследователь стоит перед выбором, какие вопросы включать в анкету и в какой последовательности их разместить. На этом этапе дорабатывается форма представления вопросов, уточняются шкалы измерения признаков, ассоциированных с вопросами. Результатом этапа является законченная форма анкеты и пробное анкетирование небольшой группы респондентов – потребителей продукта. Апробацию анкеты, или пилотное исследование, по важности можно было бы выделить и как отдельный этап.

На этом же этапе исследователь обдумывает правила организации сбора анкет, т.е. ответы на вопросы: кто, где, когда, как, у кого будет проводить опрос. При организации опросов необходимо иметь в виду вопросы репрезентативности выборок. Разрабатывая систему сбора информации, исследователь должен учитывать доступные ему ресурсы.

На пятом этапе непосредственно происходит опрос потребителей продукта и заполнение анкет на бумажном носителе.

На шестом – производится разработка базы данных для представления их в электронном виде. База данных должна разрабатываться с учетом требований к представлению данных, предъявляемых программными средствами, которые предполагается использовать для их обработки.

На седьмом этапе осуществляется ввод данных, а также оценка достоверности и подавления ошибок в данных. Этот этап, основанный на многомерном анализе выборок, уже был нами рассмотрен выше.

На восьмом этапе проводится предварительный анализ данных: исследуется структура одномерных признаков, оценивается степень зависимости пар признаков и т.п.

На девятом этапе выбирается программное средство, с помощью которого можно выполнить кластерный анализ. Если исследователь не имеет опыта применения программ классификации, это не должно останавливать его. Для выбора программного продукта достаточно знания только основ теории классификации.

Распространенные программные продукты достаточно близки по своим характеристикам. Они, как правило, предоставляют достаточно широкие возможности для эксперимента, намного превосходящие потребности пользователя, не знакомого с теорией МСА. При всем обилии вариантов и возможностей по классификации данных многомерной выборки необходимо учитывать, что если выборка действительно содержит хорошо разделяемые классы, то большинство методов дадут одни и те же или близкие результаты. С другой стороны, если изучаемая выборка вообще не содержит классов, то даже самый изощренный алгоритм их не обнаружит. Эксперимент по использованию разных метрик сходства и правил классификации может дать эффект, но слишком рассчитывать на успех в решении задач анализа анкетных данных не нужно. В случае с

потребителями туристского продукта неэффективно искать алгоритм, который дал бы лучший результат классификации, эффективнее произвести поиски признаков, которые бы лучше работали на разделение выборки, тем более, что исследователь сам их формирует.

И только **на десятом этапе** можно приступить к выявлению однородных групп объектов (наблюдений) выборки. Собственно на этом этапе и производится сегментирование опрошенных потребителей. Следует отметить, что ошибки, допущенные исследователем на любом из предшествующих этапов, могут привести к отрицательному результату при решении задачи кластеризации. В этом случае исследователь должен проанализировать все этапы с тем, чтобы понять, где именно он допустил ошибку, и повторить исследование на новом уровне именно с этого этапа.

Этап выделения сегментов потребителей рассмотрим более подробно, разбив его на подэтапы, или действия.

Первое действие. Необходимо выбрать некоторый список признаков, который, по мнению исследователя, в совокупности может описать сегменты. Если задача решается впервые, следует включать в рассмотрение избыточное количество признаков. Однако важно, чтобы признаки были измерены в одинаковых шкалах. Иногда допускается включение в список и признаков, измеренных в номинальной шкале отношений, и признаков, измеренных в ранговой шкале.

Второе действие. Если в список включены признаки, представленные в различных шкалах, необходимо произвести преобразование признаков от более богатой шкалы к более бедной. Часто при использовании признаков, заданных в шкале отношений, целесообразно произвести операцию нормировки. Преобразование ответов открытого вопроса путем типизации тоже можно отнести к этому этапу. В результате преобразования признаков будет сформирована выборка, предназначенная для классификации.

Третье действие. Воспользоваться выбранной программой для классификации сформированной выборки. Для начала достаточно выбрать небольшое количество классов (2–3) и несложный вариант оценки меры сходства (например, евклидову меру).

Четвертое действие. Производится предварительная оценка результатов классификации. Если результаты покажут, что преобладающее количество объектов объединяется в один класс, а другие классы будут включать единичные объекты (до 5% выборки), то можно сделать вывод, что классификация для данной выборки и при данном способе не удалась. В этом случае можно еще предпринять несколько попыток классификации выборки, меняя метрики сходства и другие параметры программы-классификатора, тем более что это не требует больших затрат времени. При этом можно добиться положительного результата, но маловероятно, что картина резко изменится.

Если эксперимент с параметрами программы не даст результатов, то необходимо снова обратиться к выбору нового списка признаков (первое действие) или попробовать поработать над преобразованием уже выбранных (второе действие). Указанные действия нужно повторять до тех пор, пока не будет получен правдоподобный вариант классификации.

Если и это не приведет к результату, необходимо рассмотреть процесс по этапам исследования в обратном порядке и постараться обнаружить ошибку на этих этапах или повторить исследования с первого этапа, исключив исходные гипотезы как ошибочные.

Но и в этом случае проделанная работа не бесполезна. По крайней мере, результаты предварительного анализа дают дополнительные знания об объекте, соответственно и почву для генерации новых гипотез о структуре потребителей продукта.

Если получен удовлетворительный вариант классификации, необходимо перейти к **пятому действию**: произвести более детальный анализ полученного варианта классификации (первый уровень оценки качества классификации). Во-первых, анализируются средние значения признаков по классам; во-вторых, анализируется матрица сходства классов; в-третьих, предпринимается попытка содержательной интерпретации (объяснения) полученного разбиения.

Судить о качестве разбиения позволяют некоторые простейшие приемы. Например, сравнение средних значений признаков в отдельных классах со средними значениями в целом по всей совокупности объектов. Если отличие классовых средних от общего среднего значения существенное, то это может являться признаком хорошего разбиения. Оценка существенности различий может быть выполнена с помощью t-критерия Стьюдента.

После этого можно вернуться к третьему действию и попытаться увеличить число классов. Если это действие приведет к положительному результату, то это только повысит качество конечного результата.

Шестое действие. Проводится второй уровень оценки качества классификации. Наиболее распространенные функционалы качества разбиения следующие:

1) сумма квадратов расстояний до центров классов;
2) сумма внутриклассовых расстояний между объектами;
3) суммарная внутриклассовая дисперсия;
4) среднее межклассовое расстояние.

Кроме названных функционалов качество классификации можно оценить и при помощи критерия Хотеллинга T^2 для проверки гипотезы о равенстве векторов средних для многомерной совокупности.

Перечень способов оценки качества классификации можно было бы продолжить. Однако важнейшим критерием качества остается содержательная интерпретация классификации, производимая исследователем. В конечном итоге все десять этапов, приводящих к выделению сегментов, направлены на то, чтобы подтвердить и оценить численно гипотезы исследователя о структуре потребителей, и никакая программа не сможет заменить самого исследователя.

Седьмое действие связано со вторым действием выбора списка признаков. При выборе исходного списка признаков предполагалась некоторая избыточность их количества. Для того чтобы содержательная интерпретация сегментов стала более лаконичной, желательно понизить размерность пространства признаков, исключив избыточные и неинформативные признаки.

После выполнения операции понижения размерности пространства признаков целесообразно повторить все процедуры со второго действия.

Проделав все необходимые действия и получив сегменты потребителей продукта, можно оценить их мощность и далее использовать эту информацию для выработки решений по структуре предложений продуктов, более точно отвечающей потребностям, т.е. спросу потребителей.

Приведенная выше методика применения МСА для сегментирования потребителей продукта была разработана на основании многократного решения задачи классификации первичных данных анкетных опросов потребителей услуг регионального туристского комплекса.

Литература

1. Деятельность туристских фирм в 2004 г.: стат. сб. – Владивосток: Приморский комитет гос. статистики, 2005. – 17 с.
2. Дубров А.М. Многомерные статистические методы: учебник / А.М. Дубров, В.С. Мхитарян, Л.И. Трошин. – М.: Финансы и статистика, 2000. – 352 с.
3. Костюкова О.И. Регулирование развития туристского бизнеса в регионе (на примере Приморского края): автореф. дис. … канд. экон. наук / О.И. Костюкова. – Владивосток, 1999. – 24 с.
4. Сигел Э. Практическая бизнес-статистика: пер. с англ. / Э. Сигел. – М.: Изд. дом «Вильямс», 2002. – 1056 с.
5. Сидоренко Е.В. Методы математической обработки в психологии / Е.В. Сидоренко. – СПб.: ООО «Речь», 2000. – 350 с.
6. Снапелев Ю.М. Моделирование и управление в сложных системах / Ю.М. Снапелев, В.А. Старосельский. – М.: Сов. Радио, 1974. – 264 с.
7. Сошникова Л.А. Многомерный статистический анализ в экономике: учеб. пособие / Л.А. Сошникова, В.Н. Тамашевич, Г. Уебе, М. Шеффер. – М.: ЮНИТИ, 1999. – 598 с.
8. Токарев Б.Е. Методы сбора и использования маркетинговой информации: учебно-практ. пособие / Б.Е. Токарев. – М.: Юристъ, 2003. – 254 с.
9. Щеникова Н.Б. Туризм как фактор экономического развития региона (на примере Приморского края) / Н.Б. Щеникова. – Владивосток: Изд-во ВГУЭС, 2002. – 30 с.