Программа заполнения пропусков данных в двумерных массивах

Шенцев И.А., студент 3 курса

Карапузов Д.Г., студент 3 курса

Научный руководитель Богатов Е.М.

г. Старый Оскол - 2015

Актуальность

В реальности приходится сталкиваться с ситуацией, когда некоторые из свойств одного или нескольких объектов отсутствуют – возникает ситуация данных с пропусками, что значительно осложняет математическую обработку, так как смещение основных статистических характеристик, таких как выборочное среднее или стандартное отклонение, например, возрастает прямо пропорционально числу пропусков.

Пути решения проблемы

На сегодняшний день в математической статистике существует несколько путей решения проблемы неполных данных :

1) исключение некомплектных объектов из исходной выборки;

2) метод взвешивания или метод максимального правдоподобия и EM-алгоритм;

3)методы заполнения по среднему и по регрессии;

Мы будем развивать последний из указанных подходов в направлении, называемым нами *методом самоподобия*.

Постановка задачи

Пусть имеется большой двумерный массив однородных данных со случайными пропускам в произвольных местах. Предполагается, что эти данные имеют экономический смысл, например, это могут быть ряды наблюдений за ценами на номер в отеле в каком-то городе (см. след. слайд). Здесь цены измерены в долларах, период наблюдений- 600 дней, общее число отелей -200 .

Ценовые данные с пропусками (фрагмент).

	Hotel 1	Hotel 2	Hotel 3	Hotel 4	Hotel 5	Hotel 6	Hotel 7	Hotel 8
date 1	76,0					83,0	125,0	74,0
date2	26,0	26,0				38,0	71,0	63,0
date 3	26,0	26,0	52,0		100,0	38,0	83,0	63,0
date 4	24,0	28,0	52,0		100,0	37,0	83,7	55,0
date 5	24,0	32,0	62,0		100,0	37,0	69,3	55,0
date 6	62	106,7	82,0			78,0	214,0	63,0
date 7	64,3	115,0	82,0			78,0	139,3	63,0
date 8	29,7	30,6	62,0		180,0	37,0	80,2	55,0
date 9	24,8	25,2			100,0	37,0	60,5	55,0
date 10	25,0	25,4	52,0		100,0	37,0	60,5	55,0
date 11	24,5	25,4	52,0		100,0	37,0	55,5	55,0
date 12	24,3	30,3	62,0			37,0	63,5	55,0
date 13	76,9	75,2	102,0			78,0	148,3	63,0
date 14	83,1	76,0	102,0			78,0	165,8	64,4
date 15	24,5	35,9	62,0		180,0	37,0	67,6	55,0
date 16	24,8	30,8			100,0	37,0	81,1	55,0
date 17	24,7	33,3	62,0		105,3	37,0	100,7	55,0
date 18	24,3	31,4	62,0		105,3	37,0	81,4	55,0
date 19	24,5	32,3	72,0		100,0	37,0	88,7	55,0
date 20	56,0	73,2	102,0			78,0	97,4	68,3
date 21	59,0	81,6	102,0			78,0	115,1	71,0
date 22	25,3	35,8	62,0		180,0	37,0	84,3	57,0
date 23	25,3	34,4			100,0	37,0	98,2	55,0
date 24	24,9	39,0	52,0		100,0	37,0	93,6	55,0

Постановка задачи

Пустые места в таблице образовались из-за того, что отели не всегда предоставляли информацию о ценах на свои номера.

Как видно из таблицы, можно выделить большие и малые пропуски в начале, в середине и в конце некоторых столбцов. Традиционное в таких случаях заполнение пропусков по среднему в данной ситуации, может сильно исказить реальную картину и не работает для пропусков в начале или в конце ряда наблюдений.

В этой связи актуальным является построение алгоритма, учитывающего динамику изменения цены в отелях со «сходной» ценовой политикой.

Метод самоподобия

При конструировании алгоритма восстановления данных мы будем руководствоваться «гипотезой избыточности данных». Точнее говоря, мы отталкиваемся от того, что механизм формирования цены может повторяться от отеля к отелю и от одного временного отрезка к другому. Это позволит нам подобрать нужные столбцы и строки таблицы для заполнения пропусков по регрессии.

Метод самоподобия включает в себя 2 этапа:

1.Кластеризация рядов наблюдений (поиск и выделений групп отелей с близкой ценовой политикой).

2. Восстановление пропусков в столбцах данных с использованием динамики изменения цен в других столбцах, соответствующим отелям, принадлежащих одному и тому же кластеру.

Этап №1 можно провести с использованием пакетов прикладных программ для стандартизованных данных.

Процедура заполнения пропусков.

Перейдём к процедуре непосредственного заполнения пропусков. Будем предполагать, что разбиение отелей на кластеры по принципу близости ценовой политики уже произошло. Тогда динамика изменения цен на номера в отелях одного кластера будет схожей. Для выработки стратегии восстановления данных, рассмотрим несколько характерных случаев:

Случай № 1:

1)**Пропуск приходится на понедельник-четверг**. Если пропуск приходится на понедельник, то мы берем значения цен в понедельник и делим их поочерёдно на значения цен во вторник для отелей данного кластера. Ищем среднее значение получившихся чисел. Для отеля с пропуском данных в понедельник (но известной цены на вторник) значение цен вторника умножаем на $\overline{X}$.

Пример для случая № 1:

Пример: На рис. 1 ситуация с пропуском в понедельник.
[B67]/[B68]=27.3/27.2=1.0036; [D67]/[D68]=44.4/44.1=1.0068;
Среднее значение=1.0056; [C67]=1.0056*[C68]=1.0056*32.4=32.6;

61	A	B	C	D
62	checkin	отель 1	отель 2	отель 3
63	Claster #	88	88	88
64	2011-07-29	52,4	43,2	58,0
65	2011-07-30	53,5	45,6	58,3
66	2011-07-31	25,0	31,4	44,3
67	2011-08-01	27,3		44,4
68	2011-08-02	27,2	32,4	44,1
69	2011-08-03	27,4	32,2	44,3

Рис. 1. Пример с пропуском в понедельник.

Случай № 2:

2) Если пропуск приходится на пятницу-воскресенье.

Если пропуск приходится на пятницу, то мы берем значения цен в пятницу данного кластера, делим их поочерёдно на значения цен пятницы следующей недели данного кластера. Ищем среднее значение получившихся чисел. Для отеля с пропуском данных в пятницу значение цен следующей пятницы умножаем на $\bar{X}$.

Пример для случая № 2:

Пример: На рис. 2 показана ситуация с пропуском в воскресенье.

[C45]/[C52]=22.4/21.1=1.0616; [D45]/[D52]=25.6/24.5=1.0448;

Ср. значение=1.0526; [B45]=1.0526*[B52]=1.0526*26=27.4

38	A	B	C	D
39	checkin	отель 1	отель 2	отель 3
40	**Claster #**	**79**	**79**	**79**
41	2011-07-06	24,8	23,2	24,5
42	2011-07-07	26,0	24,3	32,2
43	**2011-07-08**	**71,7**	**61,6**	**123,6**
44	**2011-07-09**	**63,6**	**54,4**	**62,2**
45	**2011-07-10**		**22,4**	**25,6**
46	2011-07-11	27,4	21,7	24,0
47	2011-07-12	26,7	22,1	23,1
48	2011-07-13	26,9	21,7	26,1
49	2011-07-14	30,2	23,8	50,3
50	**2011-07-15**	**58,3**	**53,6**	**89,3**
51	**2011-07-16**	**59,9**	**55,5**	**72,3**
52	**2011-07-17**	**26,0**	**21,1**	**24,5**

Рис. 2. Пример с пропуском в воскресенье

Случай № 3:

3) **Если пропуски идут сплошной линией, начиная с самого первого дня,** то мы берем значения цен тех дней данного кластера, на которые приходится последний пропуск. Делим их поочерёдно на значения цен следующих дней данного кластера. Ищем среднее значение $\overline{X}$ получившихся чисел. Для отеля с пропуском данных значение цен следующего дня умножаем на $\overline{X}$. Аналогично заполняем пропуски которые идут сплошной линией, начиная с последнего дня.

Пример для случая № 3:

Пример: На рис. 3 показана ситуация с сплошной линией пропусков, начинающейся с самого первого дня.

[B46] / [B47]=37.0/37.0=1; [D46] / [D47]=31.1/31.1=1;

Среднее значение=1; [C46]=1*[C47]=1*46.6=46.6

42	2011-07-07	37,0		31,1
43	**2011-07-08**	**78,0**		**67,8**
44	**2011-07-09**	**78,0**		**70,9**
45	**2011-07-10**	**37,0**		**36,7**
46	2011-07-11	37,0		31,1
47	2011-07-12	37,0	46,6	31,1
48	2011-07-13	37,0	46,2	31,0
49	2011-07-14	37,0	46,3	30,9
50	**2011-07-15**	**78,0**	**66,7**	**72,6**
51	**2011-07-16**	**78,0**	**66,7**	**72,3**
52	**2011-07-17**	**37,0**	**46,1**	**30,8**
53	2011-07-18	37,0	46,2	30,8

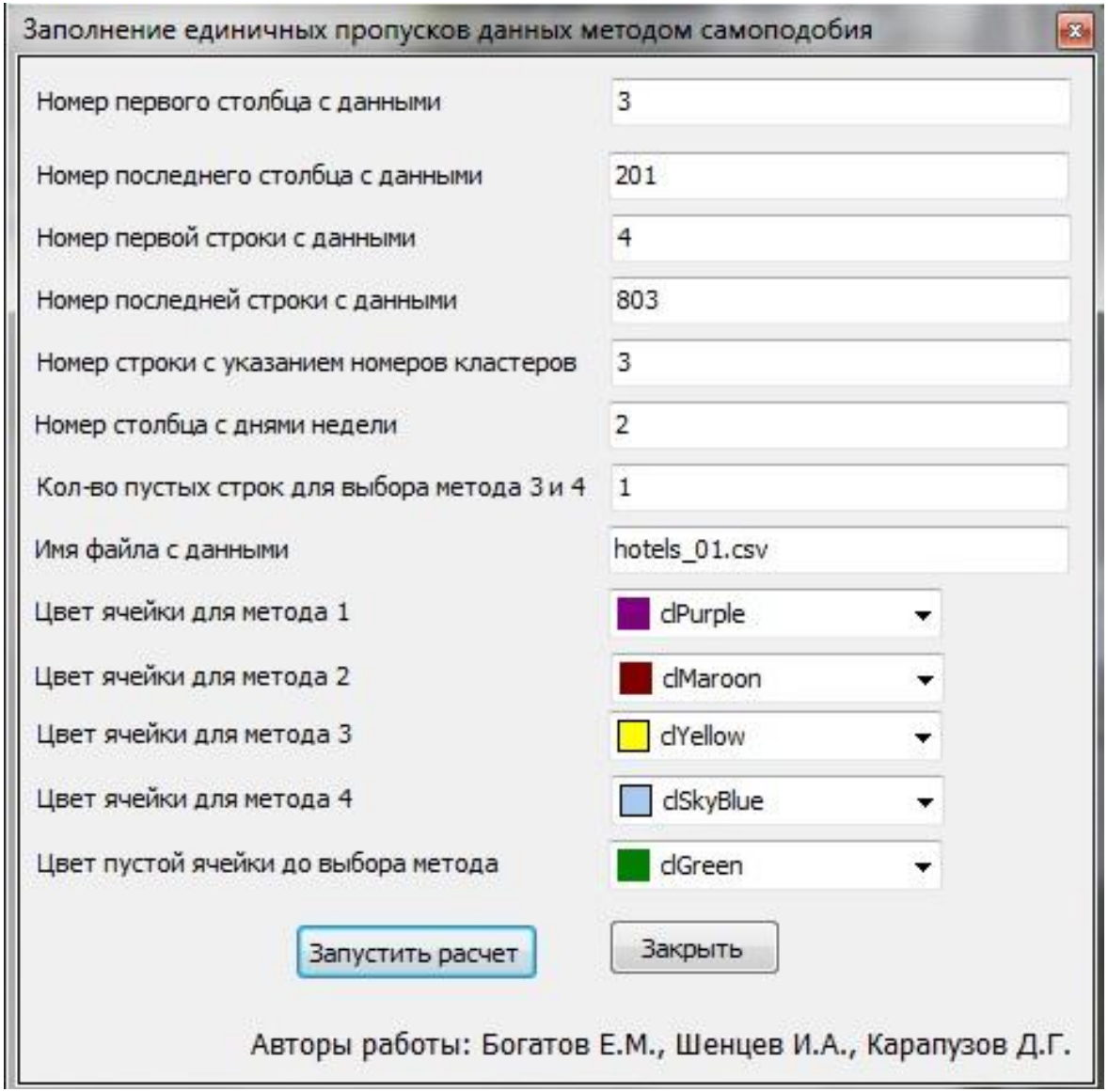
Рис. 3. Пример со сплошной линией пропусков, начинающейся с самого первого дня.

Программа заполнения пропусков.

Программный продукт предназначен для заполнения пропусков данных методом самоподобия с использованием персонального компьютера.

Для того чтобы воспользоваться программой необходимо открыть исполняемый файл «project_hotels.exe». Интерфейс программы выглядит следующим образом :

Интерфейс программы.



При нажатии кнопки «Запустить расчет» открывается заданный документ с исходными данными, затем цикл проходит по всем первым строкам и по всем последним. В случае если найдена пустая ячейка, запускается функция определения значения методом 3 или 4 (в зависимости какая строка, первая или последняя).

После этого запускается второй цикл, прохода всех ячеек, с первой до последней. В случае если обнаружена пустая ячейка, которая приходится на понедельник-четверг запускается функция определения значения методом 1, если (пятница-воскресенье) - методом 2.

Используемые технические средства

Программа написана на Embarcadero Delphi XE2. Для работы требуется современный компьютер с установленным пакетом MS Office (обязательно с установленным табличным процессором Excel). Тестирование проходило на MS Excel 2003.

Результат работы программы.

1	40,49183	173,0241	79,73701	34	47,75156	118	69,52443
3	54	78,42398	137,4762	75	63,66875	101	96,38168
3	54	79,7581	137,4762	95	63,66875	101	97,92266
5	42,0422	190,9785	168,3117	45,33397	50,935	101	74,15099
5	40,76043	181,4174	164,4902	34	50,935	118	71,89029
5	39,45679	180,0219	166,0211	28	50,935	118	69,59103
5	41,25006	174,9812	168,7179	28	50,935	118	72,75388
5	45,47209	168,9696	158,4323	33	84,89167	118	80,20039
3	56,12837	78,9957	149,2492	105	125,1883	101	114,0867
3	61,27253	82,05463	149,2492	125	125,1883	101	126,2316
3	35,81646	196,1458	143,8805	78	125,1883	101	85,93494
5	39,54677	189,007	145,0188	33	58,93607	118	84,06539
7	34	190,2812	144,2658	33	58,93607	118	75,26225
3	34	194,7487	145,931	58	58,93607	118	75,66729
2	34	198,2896	146,112	78	58,93607	118	78,33334
7	54	81,32858	149,0598	125	78,58143	101	115,9253
7	54	85,84216	156,7616	135	78,58143	101	117,7665
7	34	232,5612	134,5335	43	62,86514	101	84,71369
3	34	215,6065	134,5335	33	62,86514	118	82,497
5	34	211,6078	137,4672	33	62,86514	118	82,16702
2	34	221,3326	136,4414	33	62,86514	118	82,16702
3	34	245,8918	156,3892	38	62,86514	118	85,77862

Вывод

На основе описанного алгоритма проведена автоматизация восстановления пропусков данных в двумерных массивах.

Список литературы

- Поиск отелей [Электронный ресурс]: услуги бронирования номеров. – URLРежим доступа: http://www.hotels.com/
- Литтл, Р.Дж.А. Статистический анализ данных с пропусками / Р.Дж.А. Литтл , Д.Б. Рубин. Пер. с англ. – Москва: Финансы и статистика, 1990.- 336 с.
- Загоруйко, Н.Г. Прикладные методы анализа данных и знаний/ Н.Г. Загоруйко. – Новосибирск: Изд-во Ин-та математики, 1999. – 270 с.
- Богатов, Е.М. Кластеризация неполных данных в пакете STATISTICA/ Е.М. Богатов, В.П. Богатова // Материалы XI Всерос. научно-практической конф. с медународ. участием «Современные проблемы горно-металлургического комплекса. Наука и производство» (Старый Оскол, 3-5 декабря 2014 г.) – Старый Оскол, ОАО ОЭМК, 2014, Т.2. С. 374-378.

Спасибо за внимание.