TIME SERIES FORECASTING

ПРЕДСКАЗАНИЕ ЦЕН НА АВИАБИЛЕТЫ ПРИ ПОМОЩИ ВРЕМЕННЫХ РЯДОВ

О Б О М Н Е

А л е к с а н д р К о н д у ф о р о в

Data Science Group Leader @ **AltexSoft**

Past – .NET developer, Software Architect, PM

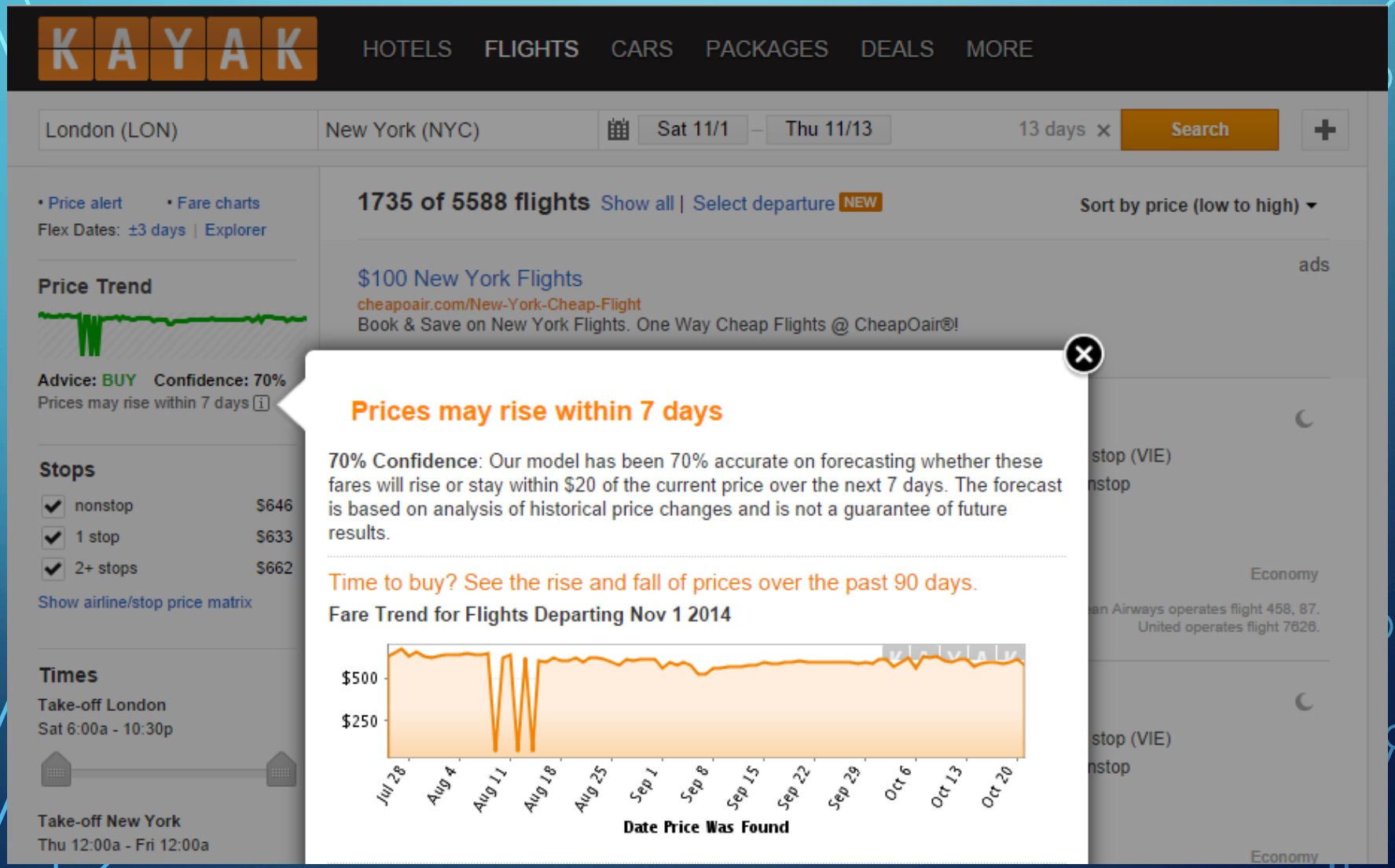
Kharkov AI club

План

1. Описание задачи прогнозирования цен на авиабилеты
2. Анализ данных
3. Временные ряды и их характеристики
4. Методы прогнозирования
5. ARIMA
6. Оценка качества прогноза, кросс-валидация
7. Описание алгоритма
8. Результаты

ПОКУПКА АВИАБИЛЕТОВ

KAYAK


KAYAK
HOTELS
FLIGHTS
CARS
PACKAGES
DEALS
MORE
London (LON)
New York (NYC)
Sat 11/1
Thu 11/13
13 days
Search
• Price alert
• Fare charts
Flex Dates: ±3 days | Explorer
Price Trend
Advice: BUY Confidence: 70%
Prices may rise within 7 days
Stops
nonstop $646
1 stop $633
2+ stops $662
Show airline/stop price matrix
Times
Take-off London
Sat 6:00a - 10:30p
Take-off New York
Thu 12:00a - Fri 12:00a
1735 of 5588 flights
Show all | Select departure NEW
Sort by price (low to high)
ads
$100 New York Flights
cheapoair.com/New-York-Cheap-Flight
Book & Save on New York Flights. One Way Cheap Flights @ CheapOair®!
Prices may rise within 7 days
70% Confidence: Our model has been 70% accurate on forecasting whether these fares will rise or stay within $20 of the current price over the next 7 days. The forecast is based on analysis of historical price changes and is not a guarantee of future results.
Time to buy? See the rise and fall of prices over the past 90 days.
Fare Trend for Flights Departing Nov 1 2014
$500
$250
Jul 28
Aug 4
Aug 11
Aug 18
Aug 25
Sep 1
Sep 8
Sep 15
Sep 22
Sep 29
Oct 6
Oct 13
Oct 20
Date Price Was Found
stop (VIE)
nstop
Economy
an Airways operates flight 458, 87.
United operates flight 7626.
stop (VIE)
nstop
Economy

ЗАДАЧА

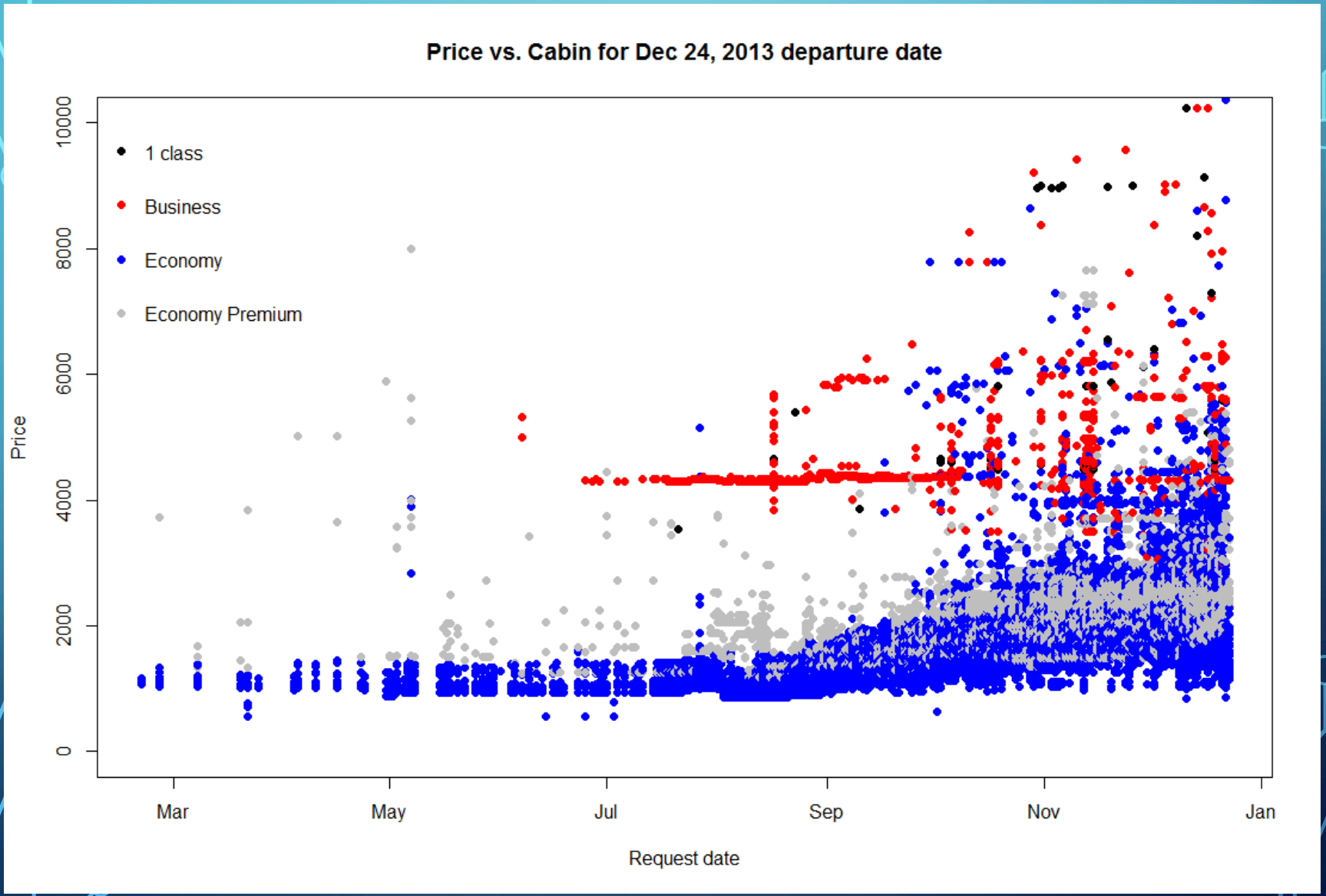
Спрогнозировать, вырастет или упадет средняя цена на авиабилеты за следующую неделю для выбранного направления и дат.

ДАННЫЕ

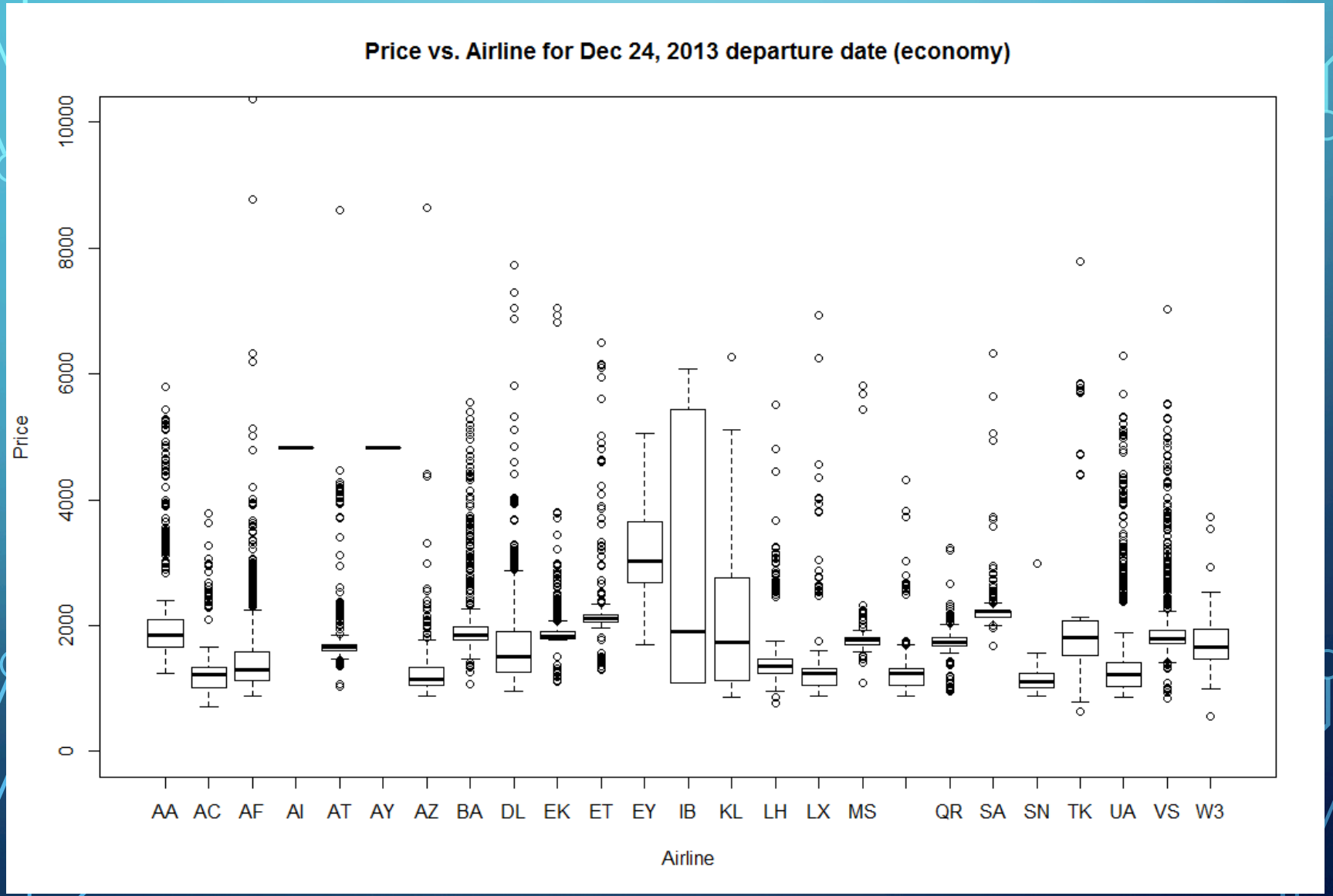
- Departure Airport/City
- Arrival Airport/City
- Departure Date
- Arrival Date
- Request date/time
- Cabin class (economy, premium economy, business, first)
- Number of Departure stops
- Number of Return stops
- Airline / Alliance
- Price amount

АНАЛИЗ ДАННЫХ

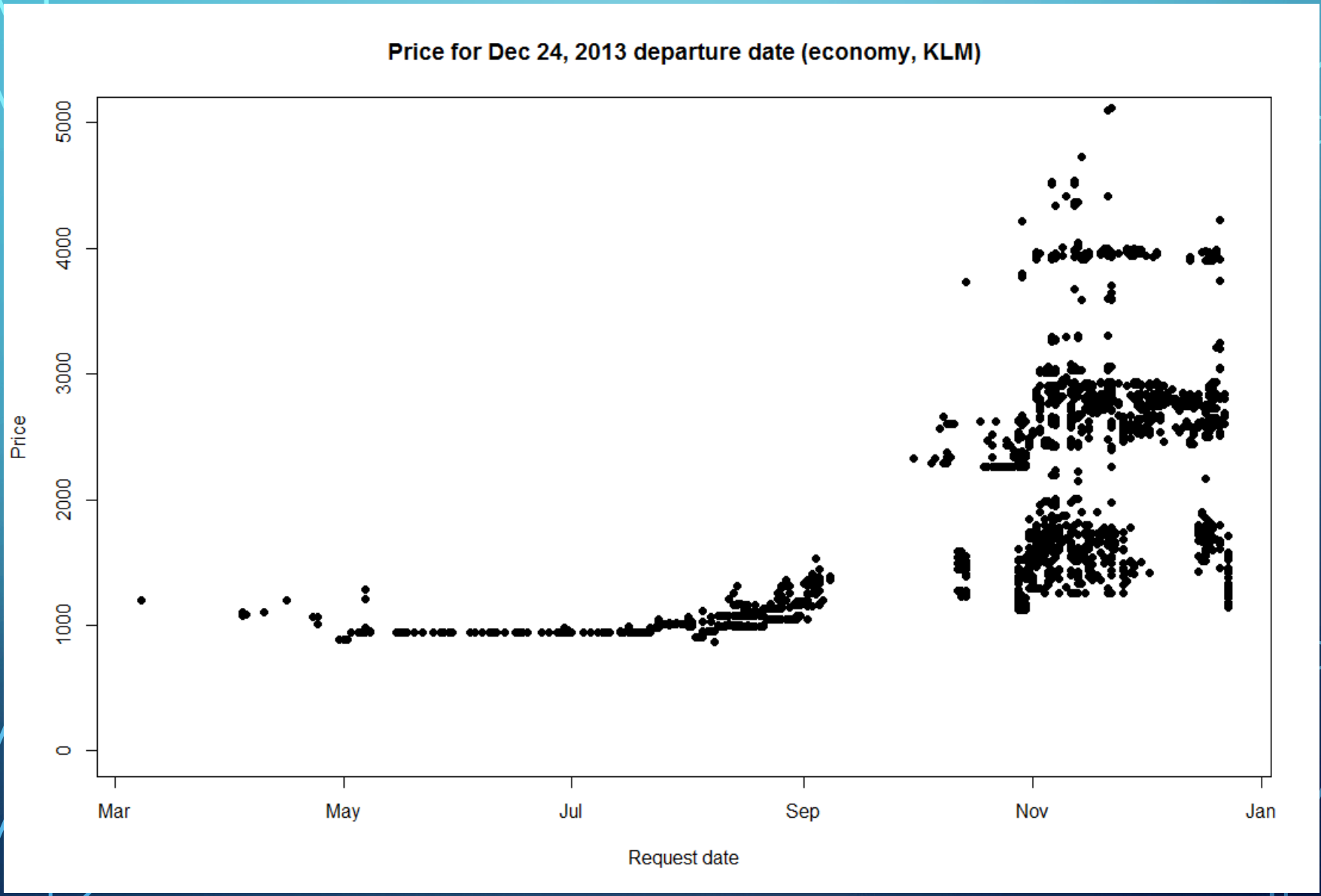
PRICE VS. CABIN



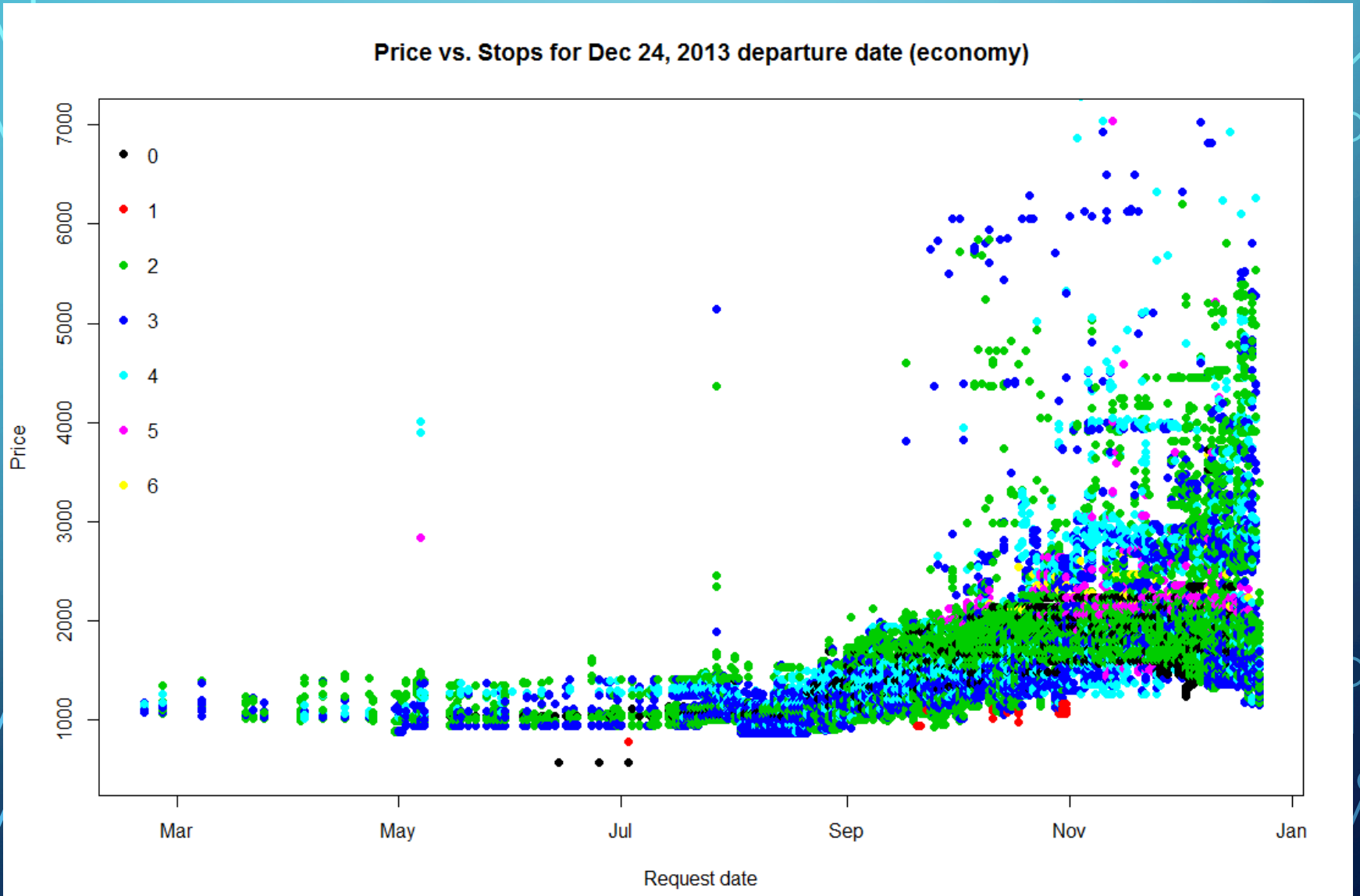
PRICE VS. AIRLINE (ECONOMY)



PRICE FOR KLM (ECONOMY)



PRICE VS. STOPS (ECONOMY)



НАБЛЮДЕНИЯ

1. Уникальные графики изменения цены для разных направлений/дат
2. Спрос на авиабилеты к праздникам выше => выше цена
3. Чем ближе к дате вылета – тем дороже цена(как правило)
4. Цена сильно зависит от класса билета(ваш КО)
5. Разные авиалинии имеют разные модели ценообразования
6. Количество остановок в пути ~ не влияет на цену
7. Внутри одной авиалинии есть

Проблема

Мы не видим многих параметров, влияющих на цену:

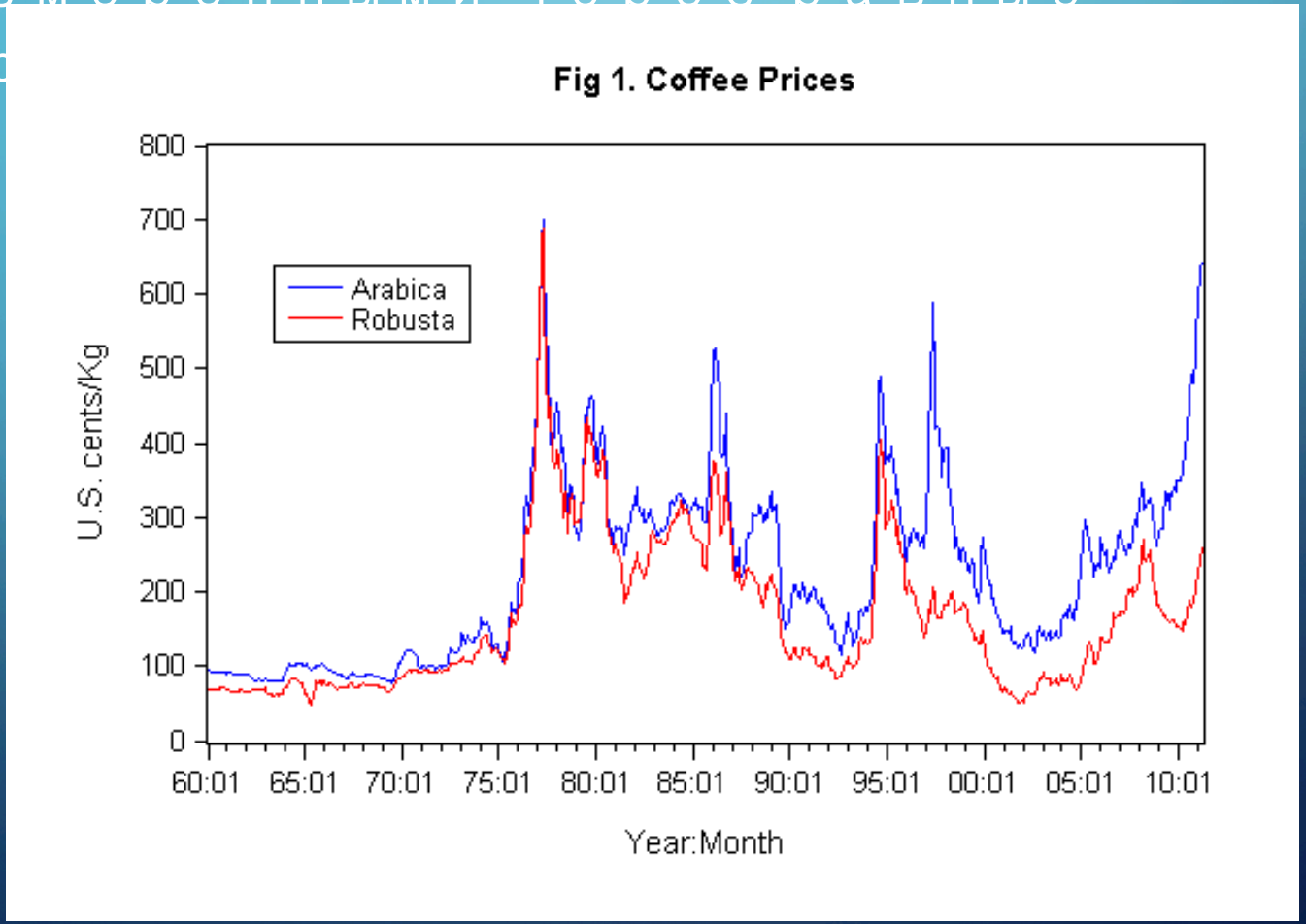
1. Количество оставшихся мест в самолете
2. Цена бензина(в разных странах)
3. Различные тарифы в аэропортах
4. Внутренние правила ценообразования авиалинии
5. Постоянная конкуренция между авиалиниями !!!

Следовательно, мы все же вынуждены работать с

ВРЕМЕННЫЕ РЯДЫ

ВРЕМЕННЫЕ РЯДЫ

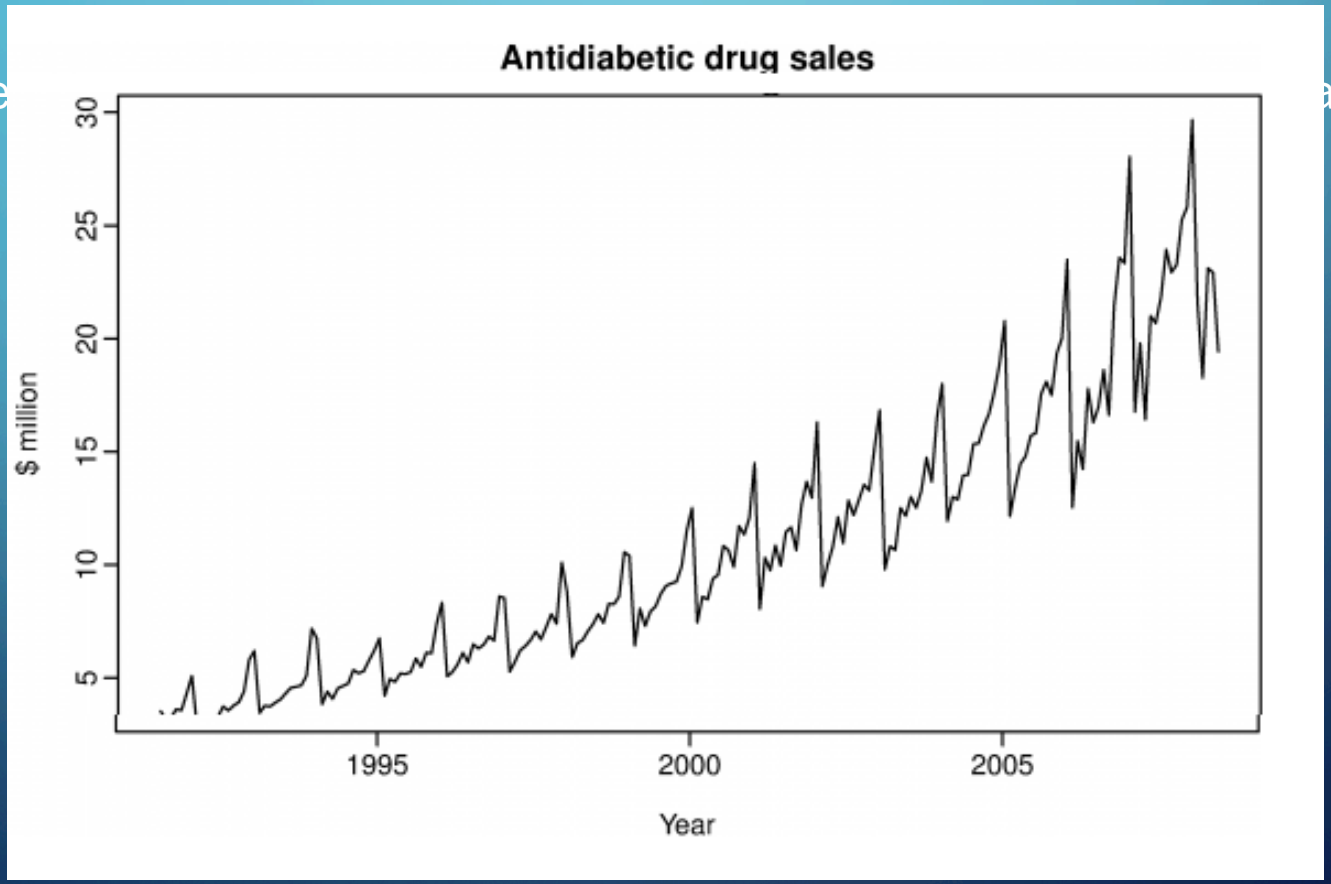
«Последовательность точек данных со значениями, измеренными через равные вр



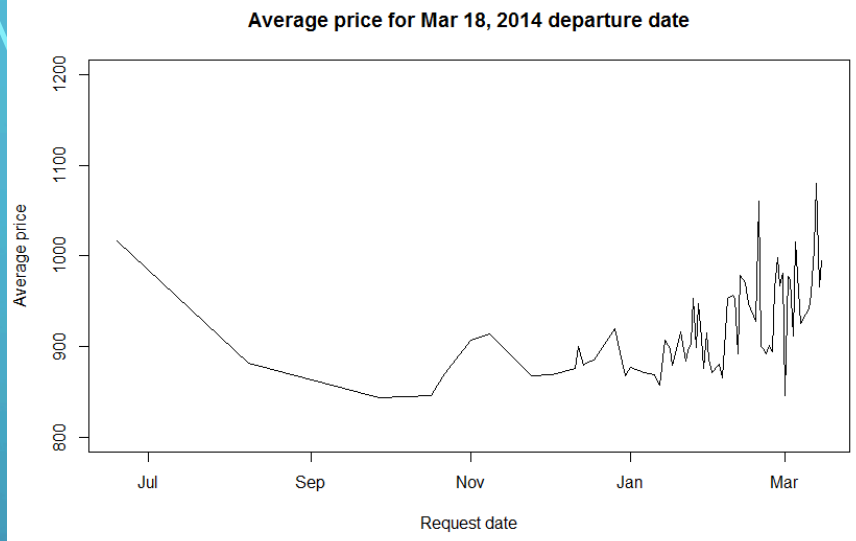
Тренд и сезонность

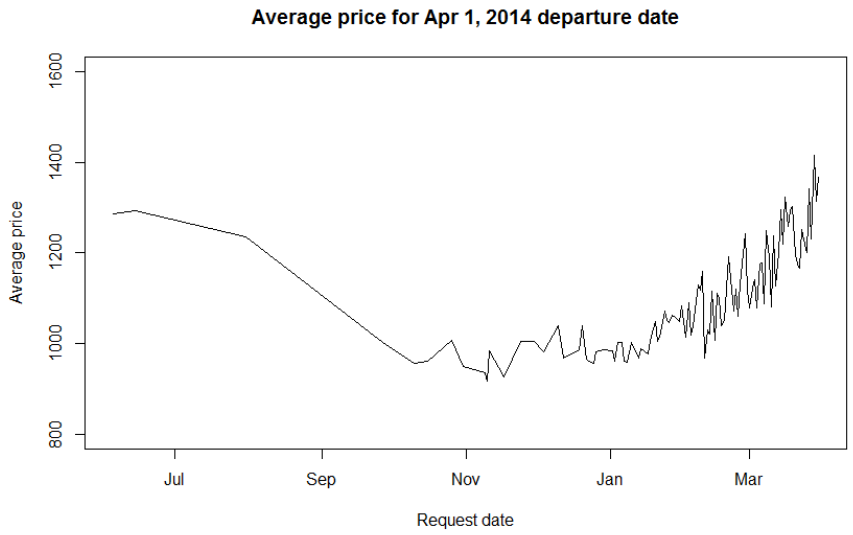
Тренд – долгосрочное увеличение или уменьшение значений ряда.

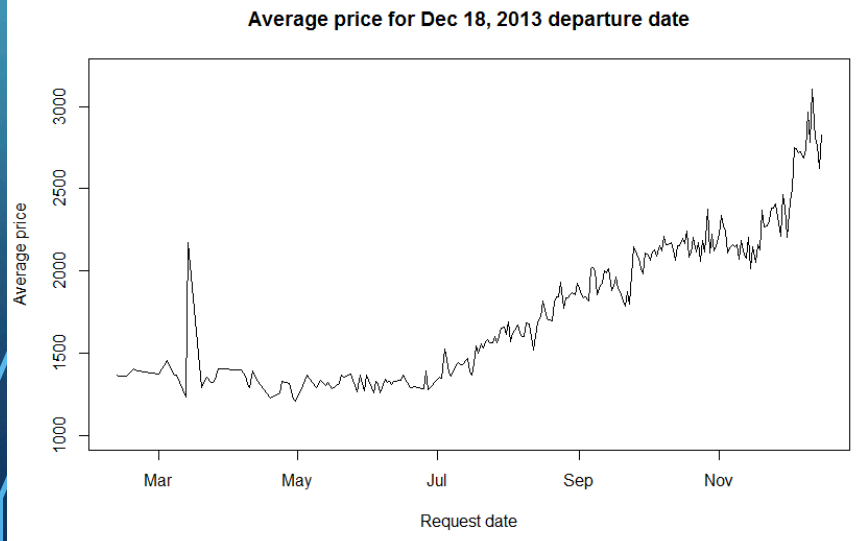
Сезонность – зависимость значений ря
«се
дн

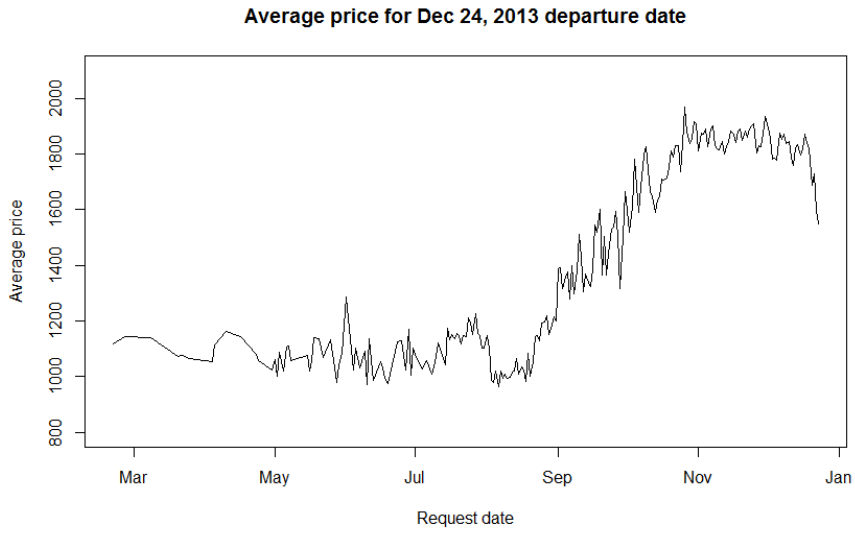


ТРЕНД И СЕЗОННОСТЬ (БИЛЕТЫ)









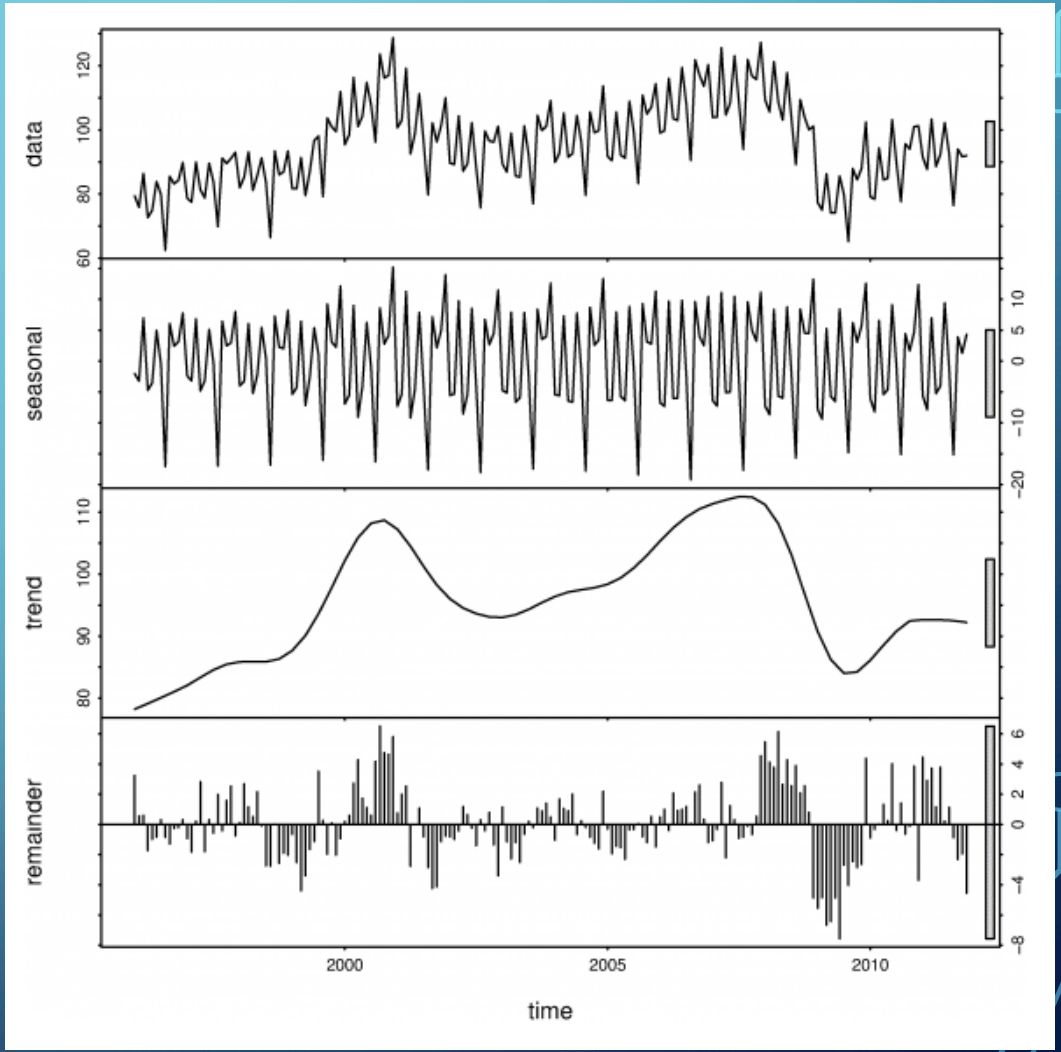
ДЕКОМПОЗИЦИЯ

Additive model:

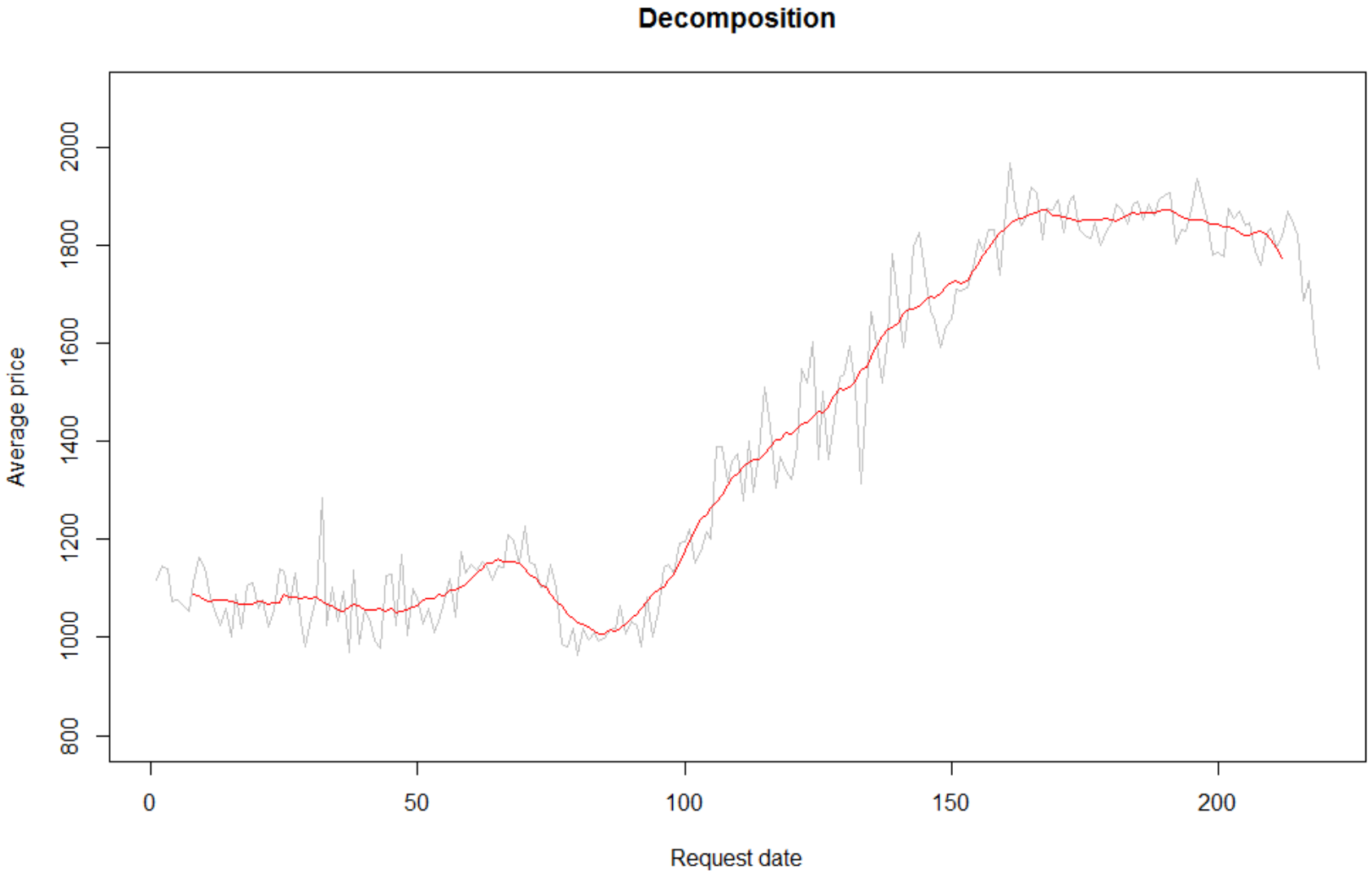
$Y_t = S_t + T_t + E_t$

Multiplicative model:

$Y_t = S_t \times T_t \times E_t$



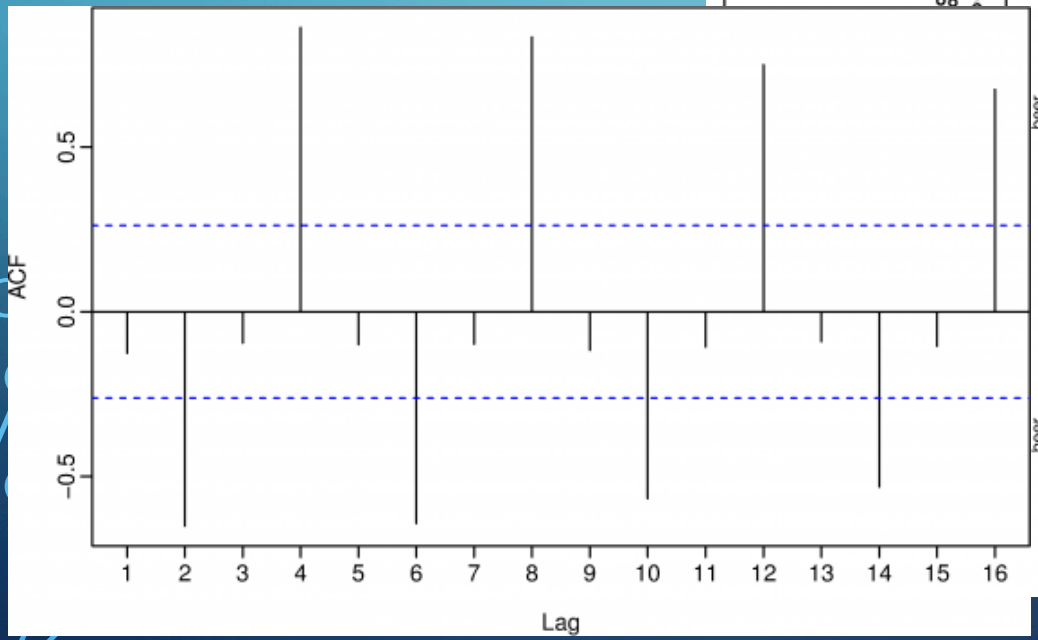
ДЕКОМПОЗИЦИЯ (АВИАБИЛЕТЫ)

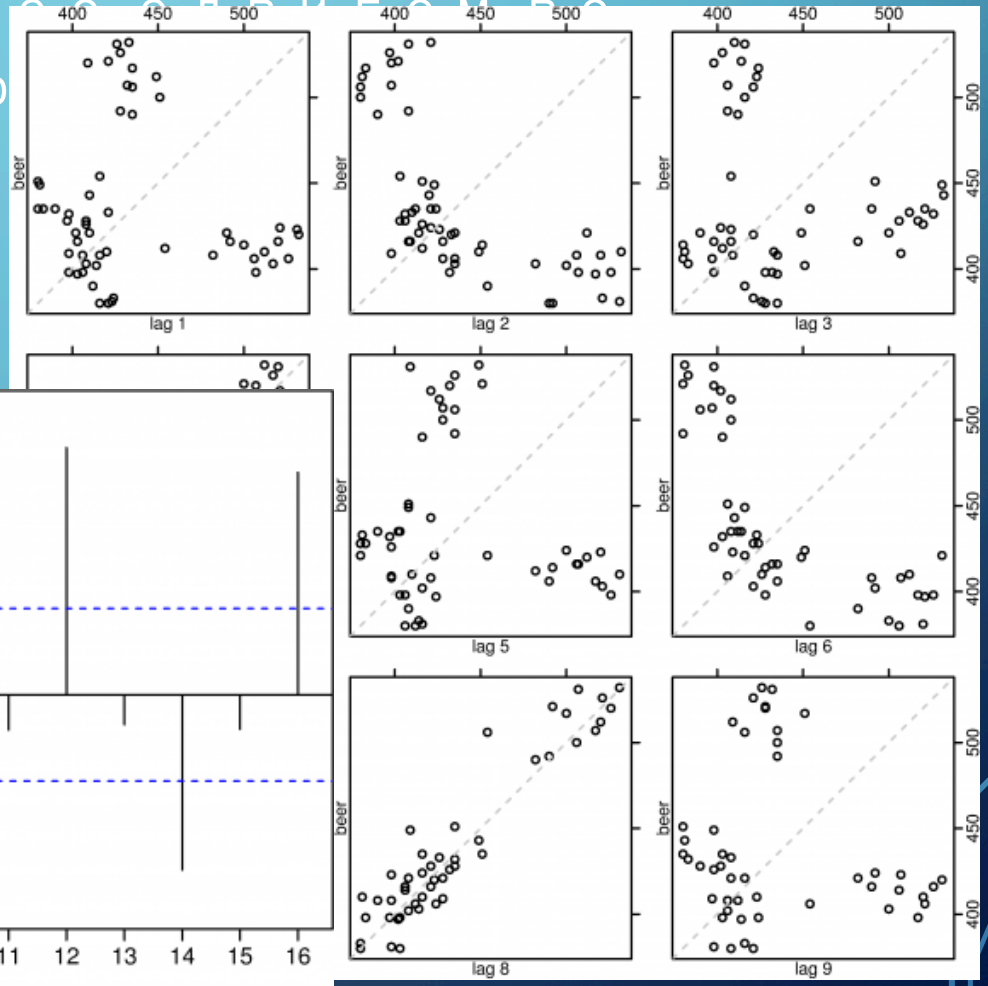


АВТОКОРРЕЛЯЦИЯ

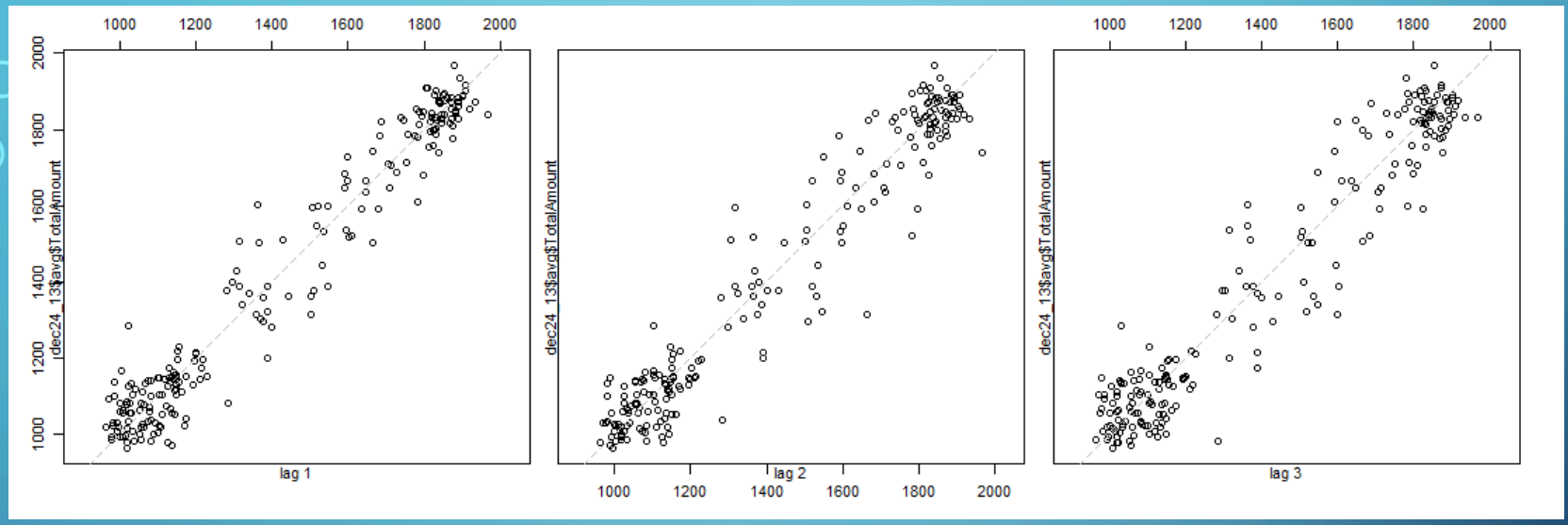
«Взаимосвязь между значениями ряда, взятыми [illegible] времени – лаго

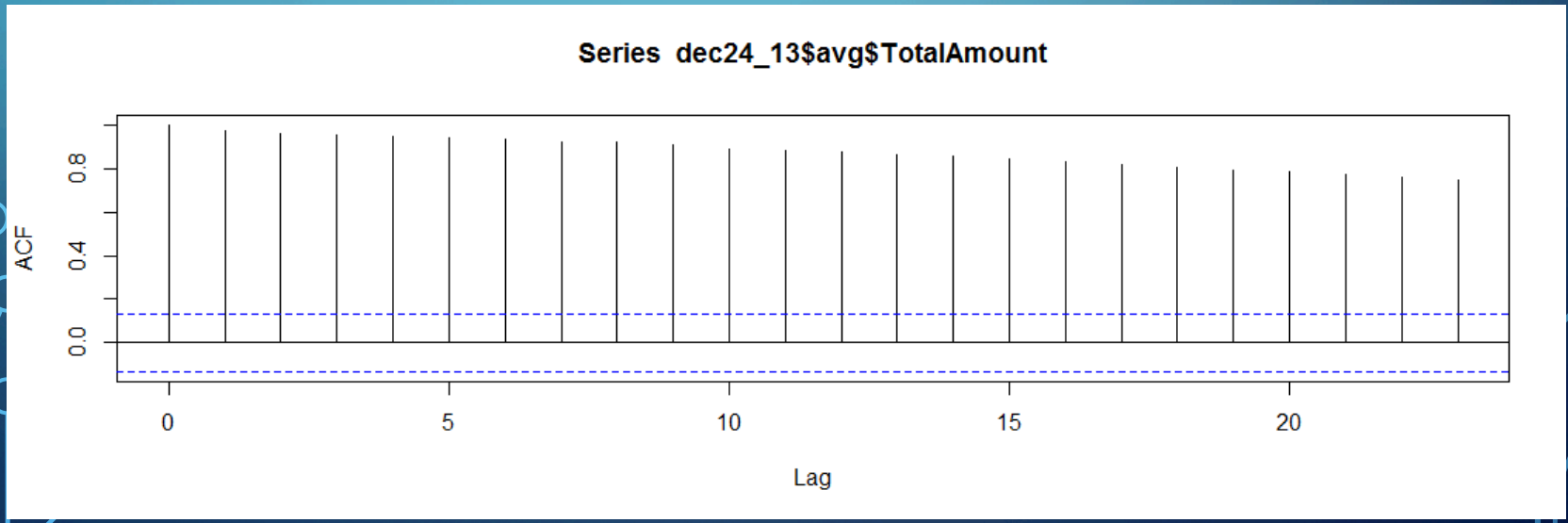
ACF





АВТОКОРРЕЛЯЦИЯ (АВИАБИЛЕТЫ)





Характеристики наших рядов

1. Нет сезонности, есть тренд => сложнее предсказывать
2. Есть ряды, в которых сильная автокорреляция, длинная или короткая => хорошо сработает авторегрессия
3. Есть ряды, в которых слабая автокорреляция
4. Некоторые ряды стационарны, некоторые нет

ПРОГНОЗИРОВАНИЕ: ТЕОРИЯ

МЕТОДЫ ПРОГНОЗИРОВАНИЯ

- Простые
  - Average (forecast = среднее значения ряда)
  - Naive (forecast = последнее значение ряда)
  - Drift (forecast = экстраполяция прямой от первого до последнего значения)
- Экспоненциальное сглаживание
- ARIMA
- Нейросети
- Динамические регрессионные модели

ARIMA

AR + I + MA (Autoregressive + Integrated + Moving Average)

ARIMA(p, d, q)

p – параметр авторегрессии (0, 1, 2, …)

d – уровень дифференцирования (0, 1, 2, …)

q – параметр скользящего среднего (0, 1, 2, …)

АВТОРЕГРЕССИЯ

Использует предыдущие p значений в линейной регрессии.

$AR(p)$: $y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + e_t$

c – константа

y_{t-p} – предыдущие значения ряда

φ_p – коэффициенты регрессии

e_t – белый шум

Скользящее среднее

Использует предыдущие q ошибок предсказания.

$MA(q)$: $y_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$

c – константа

e_{t-q} – ошибки предыдущих предсказаний

θ_p – коэффициенты

e_t – белый шум

Дифференцирование

Трансформация ряда путем вычисления разницы между рядом стоящими значениями ряда.

$$y'_t = y_t - y_{t-1}$$

Бывают разные порядки дифференцирования: d=(0,1,2,3,...)

Используется для

ПРОГНОЗИРОВАНИЕ: ПРАКТИКА

ПРОГНОЗИРОВАНИЕ

Вопрос: вырастет или упадет средняя цена билета за следующие 7 дней?

Подобрать параметры ARIMA (p, d и q) для данного ряда таким образом, чтобы прогноз был максимально точным, т.е. ошибка наименьшей.

КРОСС-ВАЛИДАЦИЯ ДЛЯ TS

Каноническая способ:

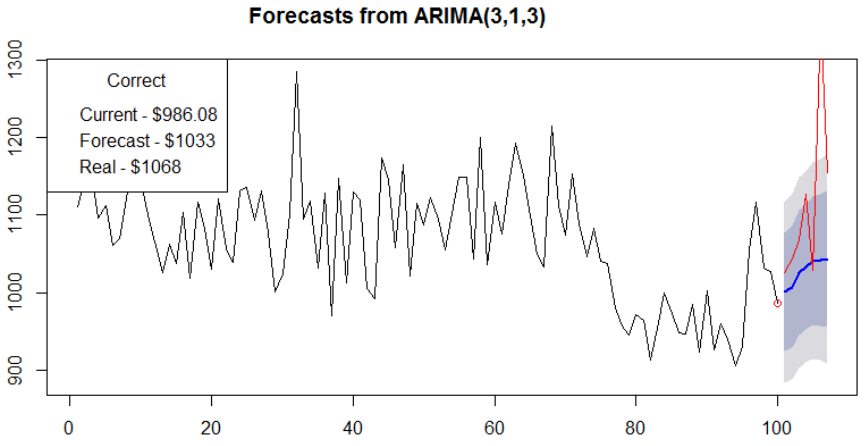
```
for (i = length(ts)-1; i > min_fit_size; i--) {
        fit ARIMA(p,d,q) model on train subset (1:i)
        forecast i+1 element
        calculate error
}
```

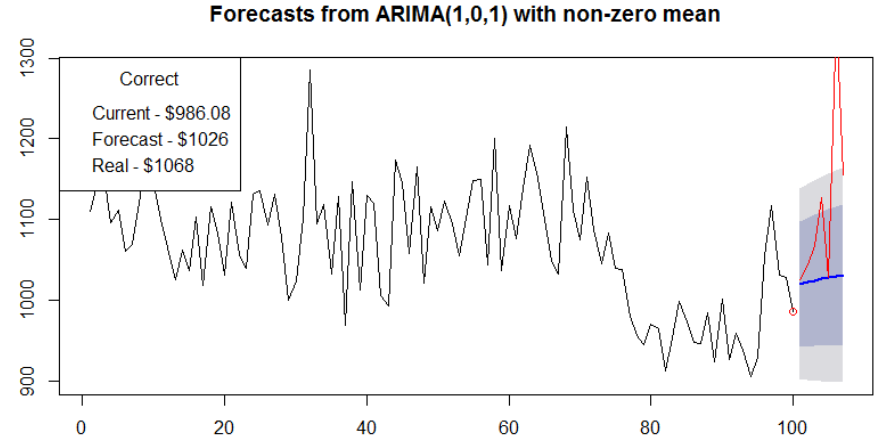
Ошибка:

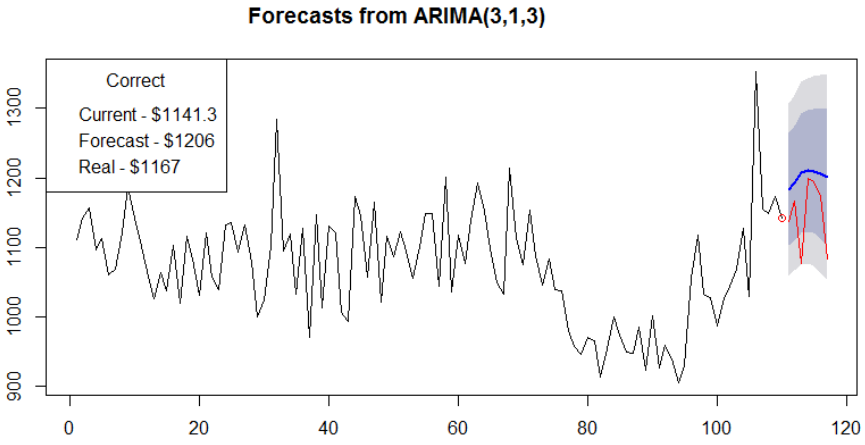
$e_i = if(sign(avg(forecast_7) - y_i) == sign(avg(valid_7) - y_i), 0, 1)$

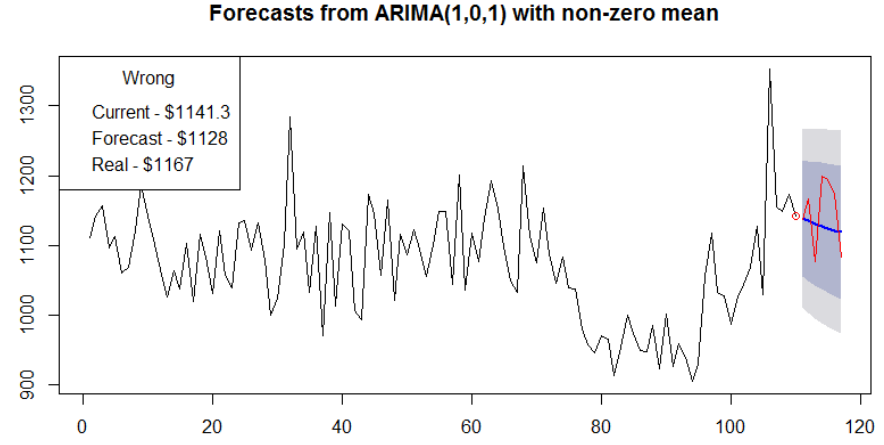
$E = \sum e_i \; / \; tests\_num$

ПРИМЕРЫ


Forecasts from ARIMA(3,1,3)
Correct
Current - $986.08
Forecast - $1033
Real - $1068
900
1000
1100
1200
1300
0
20
40
60
80
100


Forecasts from ARIMA(1,0,1) with non-zero mean
Correct
Current - $986.08
Forecast - $1026
Real - $1068
900
1000
1100
1200
1300
0
20
40
60
80
100


Forecasts from ARIMA(3,1,3)
Correct
Current - $1141.3
Forecast - $1206
Real - $1167
900
1000
1100
1200
1300
0
20
40
60
80
100
120


Forecasts from ARIMA(1,0,1) with non-zero mean
Wrong
Current - $1141.3
Forecast - $1128
Real - $1167
900
1000
1100
1200
1300
0
20
40
60
80
100
120

Алгоритм

1. Перебираем (p, d, q) от (0, 0, 0) до (p_{max}, d_{max}, q_{max})
2. Для каждого набора делаем CV и вычисляем ошибку
3. Для N моделей с наименьшей ошибкой:
   1. Делаем прогноз для каждой модели
   2. Голосование
4. Возвращаем прогноз + confidence level

РЕЗУЛЬТАТЫ

Average CV error rate = 15-20%

Average confidence = 80-85%

Некоторые ряды легче предсказать, чем другие, поэтому ошибка уникальна для каждого ряда.

ОПТИМИЗАЦИИ

Количество запусков = $p_{max} \times d_{max} \times q_{max} \times$ CV_runs

Оптимизации:

- уменьшаем длину ряда для длинных рядов – минимум 20 значений
- уменьшаем количество запусков CV – 20-40 запусков
- уменьшаем p_{max}, d_{max}, q_{max} – максимум (4,1,4)
- итеративно отбрасываем плохо работающие модели

Что почитать

- Forecasting: principles and practice, Rob J Hyndman, George Athanasopoulos
- Time Series Analysis: Forecasting and Control, George Box et. al.

СПАСИБО ЗА ВНИМАНИЕ

http://merle-amber.blogspot.com

http://aikharkov.wordpress.com

Email: alex.konduforov@altexsoft.com

Skype: alex_konduforov

Twitter: @konduforov