Course: Computer statistics

Lecture 11: Time series. Models. Forecast.

Lecturer: Oleksandr Dykhovychnyi

Лекція 11. Часові ряди. Моделі. Прогноз.

I. Модель x2

II. Розклад часового ряду на компоненти

- **II.1 Тренд**
- **II.2 Періодична компонента**
- **II.3 Розклад несезонних даних**
- **II.4 Розклад сезонних даних**

III. Приклади

- **III.1. Побудова ряду**
- **III.2 Розкладання несезонного ряду**
- **III.3. Розкладання несезонного ряду**
- **III.4. Прогнози з використанням експоненціального згладжування**

I. Модель

Часовий ряд - це послідовність числових значень, які описують явище, що протікає в часі, й виміряних в послідовні моменти часу, зазвичай через рівні проміжки. Таким чином, дані виявляються впорядкованими щодо невипадкових моментів часу, і, отже, на відміну від вибірок, які складаються з незалежних елементів, несуть в собі додаткову інформацію про внутрішні зв’язки. Ця інформація є основою для дослідження часових рядів.

Основною задачею аналізу часових рідів - це побудова адекватної математичної моделі числового ряду, її дослідження й прогнозування на підставі цієї моделі.

II. Розклад часового ряду на компоненти

Базовий підхід в аналізі часових рядів полягає у розкладі часового ряду на компоненти. Будемо позначати часовий ряд : $X(t), t=1,2,...$ Тоді,

$$X(t)=T(t)+C(s)+S(t)+R(t),$$

де

- $T(t)$ - **тренд** (детермінована компонента);
- $C(s)$ - **циклічна компонента;**
- $S(t)$ - **сезонна компонента;**
- $R(t)$ - **випадкова компонента.**

Такий розклад називають **адитивним**.
Можливий і **мультиплікативний** розклад:

$$X(t)=T(t)C(s)S(t)R(t).$$

Але шляхом логарифмування він зводиться до адитивного

$$\ln X(t)=\ln T(t)+\ln C(s)+\ln S(t)+\ln R(t).$$

Тому обмежимось лише адитивними моделями.

II.1 Тренд

Тренд – це невипадкова компонента ряду, яка показує основну тенденцію у поведінці ряду. Серед трендів, як правило, можна виділити такі найпростіші:

- $T(t)=a+bt$ - лінійний тренд;
- $T(t)=ae^{bt}$ - експоненційний тренд;
- $T(t)=a+bt+ct^2$ - квадратичний тренд.

Такі тренди можна будувати та оцінювати їхню якість методами регресійного аналізу.

II.2 Періодична компонента

Періодична компонента виявляється на підставі аналізу кореляційної функції. Аналіз кореляційної функції дозволяє знайти лаг, при якому кореляція найбільш висока, а, отже, зв'язок між поточним і попередніми рівнями часового ряду є найбільш тісним. Якщо значущим виявився тільки перший коефіцієнт кореляції (автокореляція із лагом 1), часовий ряд, швидше за все, містить тільки тренд (тенденцію). Якщо значущим виявився коефіцієнт автокореляції із лагу **n**, то ряд формують циклічні коливання з періодом **n**. Якщо коефіцієнти автокореляції не є значущими, то можна сказати, що або ряд не містить явного тренду і періодичних компонент, або ряд містить нелінійну тенденцію, яку лінійний коефіцієнт кореляції виявити не здатний.

II.3 Розклад несезонних даних

Несезонний часовий ряд складається тренду і випадкової компоненти. Щоб оцінити тренд несезонного часового ряду, який може бути описаний з використанням адитивної моделі, зазвичай використовується метод згладжування, приміром, простого **ковзного середнього** часового ряду. Це застосовується у випадку «хаотичних» рядів, які не мають явного тренду. **Ковзне середнє** порядку m - це наступне перетворення ряду:

$$X(t) = \frac{X(t) + X(t-1) + \ldots + X(t-m+1)}{m} .$$

Іншим способом згладжування часового ряду є експоненціальне згладжування. Воно буває трьох рівнів: просте, подвійне й потрійне. Найпростіше виглядає так:

$$X(t) = \alpha X(t) + (1-\alpha) X(t-1), \alpha \in (0,1) .$$

Більш складним є експоненціальне згладжування. Для цього розбивають ряд на дві складові - $L(t)$ - рівень (**level, intercept**) і $T(t)$ - тренд (**trend, slope**).

Тоді,

$$\begin{aligned} &L(t) = \alpha X(t) + (1-\alpha)\big(L(t-1) + T(t-1)\big), \\ &T(t) = \beta\big(L(t) - L(t-1)\big) + (1-\beta) T(t-1), \\ &\hat{X}(t) = L(t) + T(t), \end{aligned}$$

де α, β - задані сталі.

II.4 Розкладання сезонних даних

Далі для сезонних рядів обчислюється **сезонна складова**, як середнє (для адитивних моделей) або урізане середнє (для мультиплікативних моделей) всіх значень ряду, відповідних даній точці сезонного інтервалу за аналогічним часовими періодами. Якщо часовий ряд представлений адитивної моделлю, то в якості сезонної компоненти (складової) використовується показник абсолютного відхилення. Сума всіх сезонних компонент, тобто, показників абсолютних відхилень повинна дорівнювати нулю.

III. Приклади

Розглянемо краще на прикладах, як працюють деякі процедури дослідження й прогнозування часового ряду.

III.1. Побудова ряду

x2

Приклади:

```
# зчитуємо дані про тривалість життя 42 англійських королів
> kings <- scan("http://robjhyndman.com/tsdldata/misc/kings.dat",skip=3
)
Read 42 items
```

Часові ряди в **R** є спеціальним об’єктом типу **ts**(**time series**).

```
> kingsts <- ts(kings)
> kingsts # ряд має атрибути
Time Series:
Start = 1
End = 42
Frequency = 1
 [1] 60 43 67 50 56 42 50 65 68 43 65 34 47 34 49 41 13 35 53 56 16 43 69 59 4
8
[26] 59 86 55 68 51 33 49 67 77 81 67 71 81 68 70 77 56
> class(kingsts)
[1] "ts"
```

Іноді часовий ряд формується через регулярні проміжки часу, які становлять менше одного року, наприклад, щомісяця або щокварталу. В цьому випадку можна вказати кількість разів, коли дані збиралися за рік, за допомогою параметра частоти(**frequency**) в функції **ts ()**. Для даних часових рядів за місяць встановлють **frequency** = 12, а для часових рядів за квартал - **frequency** = 4. Також можна вказати перший рік, коли були зібрані дані й перший інтервал в цьому році, використовуючи опцію **start** в функції **ts ()**. Наприклад, якщо перша точка даних відповідає другому кварталу 1986 року, встановлють **start** = c (1986,2).

Приклади:

```
# зчитуємо дані про кількість новонароджених у НЙ
> births <- scan("http://robjhyndman.com/tsdldata/data/nybirths.dat")
Read 168 items
# створюємо ряд по роках і місяцях, починаючи з першого кварталу
# 1946
> birthstimeseries <- ts(births, frequency=12, start=c(1946,1))
> birthstimeseries

        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
1946 26.663 23.598 26.931 24.740 25.806 24.364 24.477 23.901 23.175 23.227 21.672 21.870
1947 21.439 21.089 23.709 21.669 21.752 20.761 23.479 23.824 23.105 23.110 21.759 22.073
.....................................................................................
```

Після того, як часовий ряд в R прочитано, наступним кроком є побудова графіка, що можна зробити за допомогою функції **plot.ts ()**. Наприклад, для побудови часового ряду кількість новонароджених у НЙ (Рис.1):

Приклади:

```
> plot.ts(birthstimeseries)
```

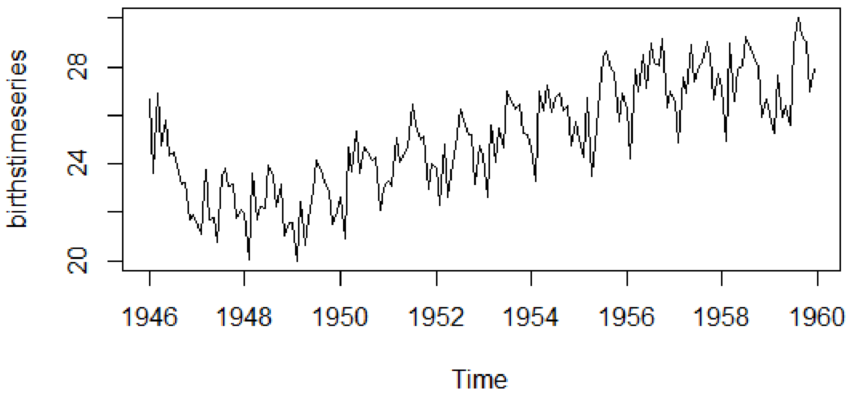

Рис. 1(Рисунок згенеровано RStudio).

З графіка (Рис.1) видно, що кількість народжень на місяць є сезонним: пік-щоліта, а западина - щозими. Цей часовий ряд, ймовірно, можна описати з використанням адитивної моделі, оскільки сезонні коливання приблизно постійні за розміром в часі і, схоже, не залежать від рівня часового ряду, а випадкові коливання також мають приблизно сталу амплітуду з плином часу.

III.2 Розкладання несезонного ряду

Розглянемо розкладання несезонного часового ряду. Такий ряд складає ться тренду й нерегулярної компоненти. Щоб оцінити компонент тренду н есезонного часового ряду, який може бути описаний з використанням адит ивної моделі, зазвичай використовується метод згладжування, такий як об числення простого ковзного середнього часового ряду. Це можна зробити за допомогою функції **SMA** пакету **TTR.** Зробимо це на прикладі ряду тривалості життя королів.

Будуємо його графік (Рис.2):

Приклади:
> plot.ts(kingsts)

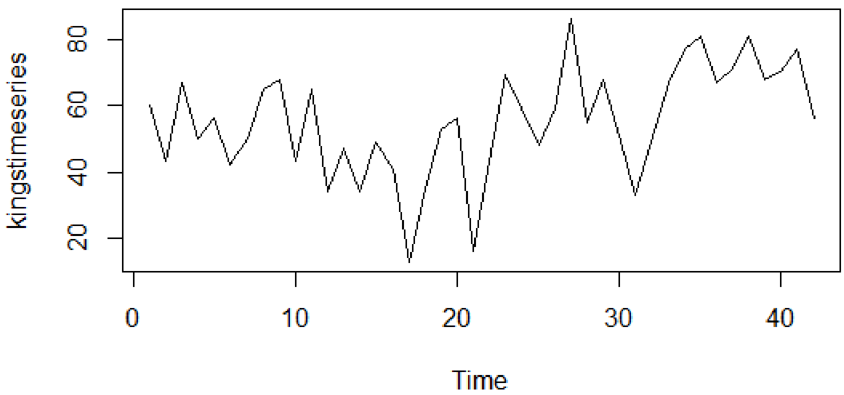

Рис.2 (Рисунок згенеровано RStudio).

Приклади:
> kingstsSMA3 <- SMA(kingsts,n=3)
Робимо згладжування методом ковзного середнього порядку 3 (Рис. 3)
> plot.ts(kingstsSMA3)

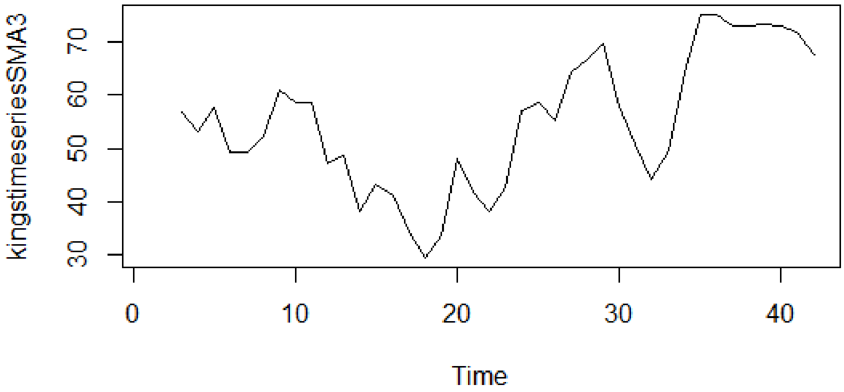

Рис.3 (Рисунок згенеровано RStudio).

Але кращий ефект дає згладжування порядку 8 (рис.4), воно дозволяє бачити тенденції у поведінці ряду.

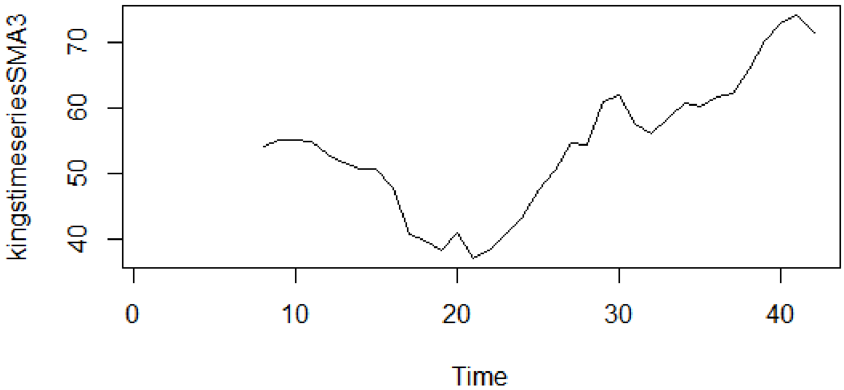

Рис.4 (Рисунок згенеровано RStudio).

Дані, згладжені за допомогою простого ковзного середнього порядку 8 (рис.4), дають більш чітку картину компоненти тренду, тому можна бачити , що вік смерті англійських королів, мабуть, зменшився з 55 років до 38 рок ів під час правління з перших 20 королів, а потім збільшилася до 73 років до кінця правління 40-го короля в часі ряду.

III.3. Розкладання несезонного ряду

Сезонний часовий ряд складається з таких компонент: тренда, сезонної компоненти й випадкової компоненти. Розкладання часового ряду означає поділ часового ряду на ці три компоненти.

Часовий ряд числа народжень в місяць в місті Нью-Йорку є сезонним з піком щоліта і проваллям щозими і може бути описаний з використанням а дитивної моделі, оскільки сезонні і випадкові коливання приблизно постійні за розміром у часі. Розклад на ці компоненти реалізує функція

decompose(…),

яка повертає їх як елементи списку.

Приклади:

```
> birthstimeseriescomponents <- decompose(birthstimeseries) # розклад
> birthstimeseriescomponents$seasonal #сезонна копонента
            Jan        Feb        Mar        Apr        May        Jun
1946 -0.6771947 -2.0829607  0.8625232 -0.8016787  0.2516514 -0.1532556
1947 -0.6771947 -2.0829607  0.8625232 -0.8016787  0.2516514 -0.1532556
            Jul        Aug        Sep        Oct        Nov        Dec
1946  1.4560457  1.1645938  0.6916162  0.7752444 -1.1097652 -0.3768197
1947  1.4560457  1.1645938  0.6916162  0.7752444 -1.1097652 -0.3768197
```

Будуємо графік тренда, сезонної і випадкової компонент, використ овуючи функцію **plot ()** (Рис.5):

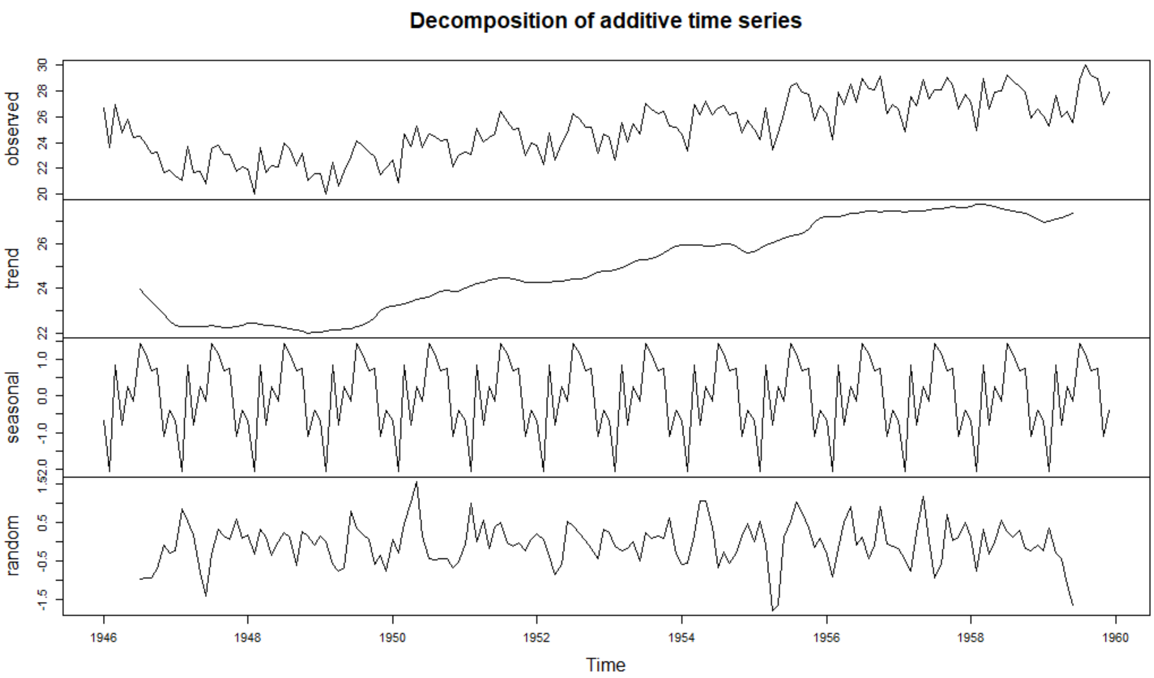

Рис.5(Рисунок згенеровано RStudio).

III.4. Прогнози з використанням експоненціального згладжування

Експоненціальне згладжування може використовуватися для створення короткострокових прогнозів для даних часових рядів. Покажемо це на прикладі ряду, який складає загальна річна кількість опадів у дюймах для Лондона за 1813-1912 рр (Рис.6).

Приклади:

```
> rain <- scan("http://robjhyndman.com/tsdldata/hurst/precip1.dat",skip=1)
Read 100 items
> rainseries <- ts(rain,start=c(1813))
> plot.ts(rainseries)
```

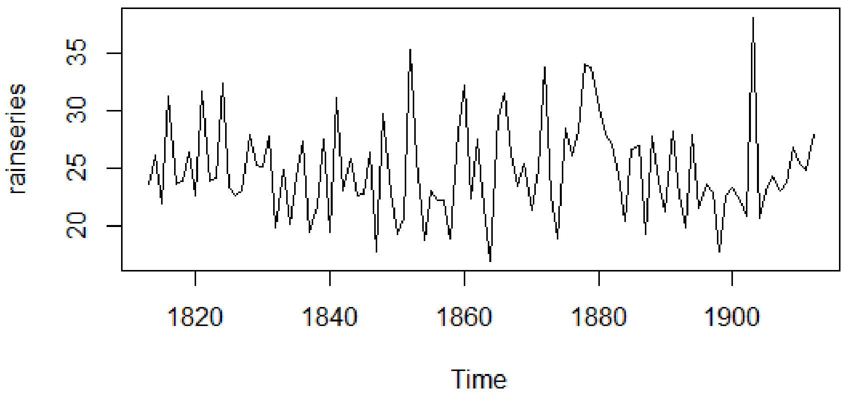

Рис.6 (Рисунок згенеровано RStudio).

Як видно рисунку 6, випадкові коливання схожі на приблизно постійними за розміром у часі, тому, ймовірно, доцільно описати дані з використанням адитивної моделі. Таким чином, побудуємо прогноз, використовуючи просте експоненціальне згладжування. Щоб робити прогнози з використанням простого експоненціального згладжування в R, будується проста модель експоненціального згладжування з використанням функції **HoltWinters ()**з параметрами **beta = FALSE** і **gamma = FALSE**.

Приклади:

```
> rainseriesforecasts <- HoltWinters(rainseries, beta=FALSE, gamma=FALSE)
> rainseriesforecasts

Holt-Winters exponential smoothing without trend and without seasonal component.

Call:
```

HoltWinters(x = rainseries, beta = FALSE, gamma = FALSE)

Smoothing parameters:
alpha: 0.02412151
beta : FALSE
gamma: FALSE

Вихідні дані **HoltWinters ()** показують, що передбачуване значення альфа-параметра становить близько 0,024. Це дуже близько до нуля, що свідчить про те, що прогнози засновані як на недавніх, так і на менш недавніх спостереженнях (хоча деякі недавні спостереження мають більшу вагу). Побудуємо графік ряду разом з прогнозом(червоний колір) (Рис. 7)

Приклади:
>plot(rainseriesforecasts)

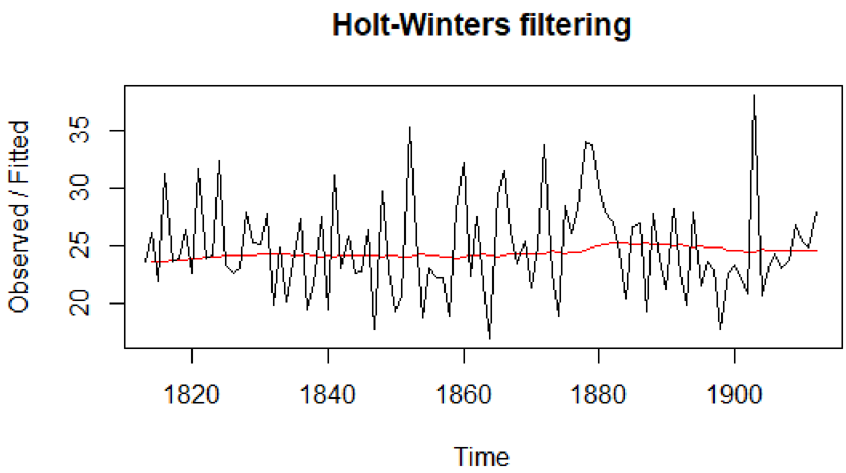

Рис.7 (Рисунок згенеровано RStudio).

Зпрогнозуємо цей ряд на вісім кроків уперед за допомогою функції **forecast().**

Приклади:
> rainseriesforecasts2 <- forecast(rainseriesforecasts, h=8)
> rainseriesforecasts2

```
     Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
1913       24.67819 19.17493 30.18145 16.26169 33.09470
1914       24.67819 19.17333 30.18305 16.25924 33.09715
1915       24.67819 19.17173 30.18465 16.25679 33.09960
1916       24.67819 19.17013 30.18625 16.25434 33.10204
1917       24.67819 19.16853 30.18785 16.25190 33.10449
1918       24.67819 19.16694 30.18945 16.24945 33.10694
1919       24.67819 19.16534 30.19105 16.24701 33.10938
1920       24.67819 19.16374 30.19265 16.24456 33.11182
```

Функція **forecast ()** надає прогноз на рік, інтервал прогнозування 80% дл я прогнозу й інтервал прогнозування 95% для прогнозу. Наприклад, прогно зована кількість опадів в 1920 році становить близько 24,68 дюймів з інтер валом прогнозування 95% (16,24, 33,11). При такому простому згладжуван ні прогноз є сталою 24.67819.

Для побудови графіків прогнозів, зроблених функцією **forecast ()** викор истовуємо функцію **plot** (Рис.8):

Приклади:

>plot(rainseriesforecasts2)

(Рисунок згенеровано RStudio).

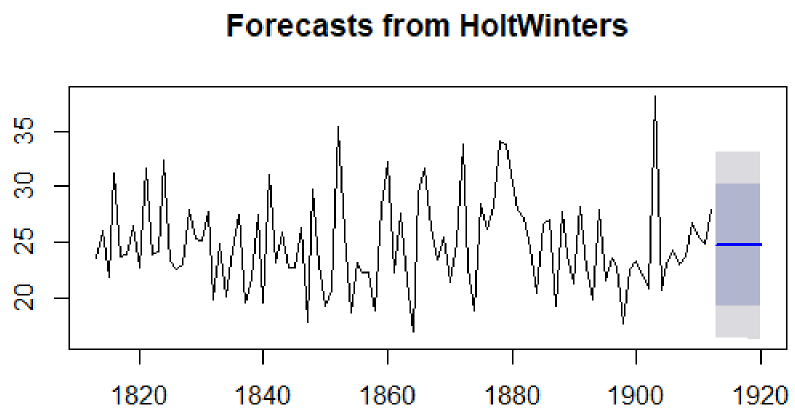

Рис.8 (Рисунок згенеровано RStudio).

Тут прогнози на 1913-1920 рр. зображені у вигляді синьої лінії, інтерва л прогнозування 80% - заштрихована область фіолетового кольору, а інтервал прогнозування 95% - як заштрихована сіра область.