НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ   
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»  
ФАКУЛЬТЕТ ІНФОРМАТИКИ ТА ОБЧИСЛЮВАЛЬНОЇ ТЕХНІКИ  
Кафедра автоматизованих систем обробки інформації та управління

**ЗВІТ**

до лабораторної роботи №1

з дисципліни «Інтелектуальний аналіз даних»

Виконав:

студент гр. ІС-32

Капорін Р. М.

Київ – 2016

**Мета:** спрогнозувати погоду в Житомирі за 2013 рік.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ЯНВ | ФЕВ | МАР | АПР | МАЙ | ИЮН | ИЮЛ | АВГ | СЕН | ОКТ | НОЯ | ДЕК |
| 1 | 5 | 2 | 1 | 2 | 15 | 19 | 20 | 25 | 23 | 8 | 15 | 3 |
| 2 | 3 | 1 | 0 | 4 | 20 | 19 | 24 | 26 | 19 | 6 | 15 | 3 |
| 3 | 2 | 1 | 1 | 3 | 28 | 17 | 25 | 23 | 14 | 6 | 18 | 1 |
| 4 | 2 | -1 | -2 | 1 | 18 | 22 | 27 | 28 | 19 | 9 | 13 | 2 |
| 5 | 1 | 4 | 4 | 4 | 22 | 25 | 29 | 30 | 19 | 10 | 16 | 2 |
| 6 | 0 | 4 | 9 | 2 | 23 | 27 | 30 | 30 | 19 | 15 | 14 | 1 |
| 7 | -5 | 3 | 5 | 2 | 26 | 26 | 26 | 31 | 21 | 16 | 9 | 0 |
| 8 | -6 | -1 | 1 | 6 | 27 | 27 | 24 | 32 | 20 | 16 | 16 | -2 |
| 9 | -8 | -1 | -3 | 8 | 28 | 27 | 28 | 33 | 23 | 13 | 14 | 2 |
| 10 | -3 | -2 | -1 | 12 | 27 | 28 | 27 | 32 | 18 | 11 | 12 | -4 |
| 11 | -1 | 0 | -2 | 12 | 29 | 24 | 30 | 21 | 22 | 16 | 7 | 1 |
| 12 | -3 | -2 | 0 | 13 | 28 | 20 | 22 | 27 | 21 | 13 | 6 | 1 |
| 13 | -3 | 1 | 3 | 10 | 24 | 24 | 21 | 31 | 16 | 15 | 5 | 1 |
| 14 | -7 | 1 | 1 | 11 | 17 | 26 | 22 | 22 | 14 | 13 | 5 | 0 |
| 15 | 1 | -1 | -1 | 10 | 22 | 26 | 21 | 23 | 14 | 14 | 5 | 0 |
| 16 | 1 | -2 | -6 | 17 | 25 | 27 | 22 | 24 | 18 | 13 | 4 | 1 |
| 17 | 0 | -3 | -4 | 18 | 27 | 26 | 23 | 26 | 22 | 14 | 5 | 4 |
| 18 | -4 | -4 | -1 | 20 | 27 | 26 | 27 | 28 | 16 | 14 | 7 | 1 |
| 19 | -5 | -2 | 1 | 24 | 25 | 24 | 28 | 31 | 12 | 8 | 7 | -1 |
| 20 | -3 | -1 | 2 | 20 | 29 | 27 | 20 | 30 | 11 | 17 | 12 | 0 |
| 21 | -2 | -1 | 3 | 14 | 23 | 29 | 20 | 31 | 14 | 18 | 11 | 2 |
| 22 | -4 | 0 | 0 | 17 | 21 | 30 | 21 | 20 | 16 | 15 | 11 | 5 |
| 23 | -7 | 1 | -8 | 22 | 21 | 33 | 19 | 21 | 17 | 16 | 9 | 7 |
| 24 | -6 | -1 | -4 | 19 | 18 | 26 | 21 | 21 | 14 | 21 | 9 | 7 |
| 25 | -8 | 2 | -4 | 22 | 19 | 29 | 24 | 19 | 13 | 14 | 9 | 2 |
| 26 | -10 | 2 | -2 | 26 | 21 | 24 | 24 | 17 | 7 | 15 | -1 | 4 |
| 27 | -12 | 4 | -1 | 30 | 15 | 20 | 27 | 18 | 10 | 22 | -4 | 4 |
| 28 | -8 | 3 | 0 | 25 | 19 | 25 | 31 | 21 | 11 | 18 | 1 | 4 |
| 29 | -1 |  | 2 | 24 | 23 | 21 | 32 | 18 | 11 | 22 | 2 | 11 |
| 30 | 1 |  | 6 | 24 | 25 | 23 | 32 | 18 | 12 | 13 | 4 | 4 |
| 31 | 3 |  | 2 |  | 20 |  | 24 | 23 |  | 12 |  | 1 |

Рис. 1 - Початкові дані

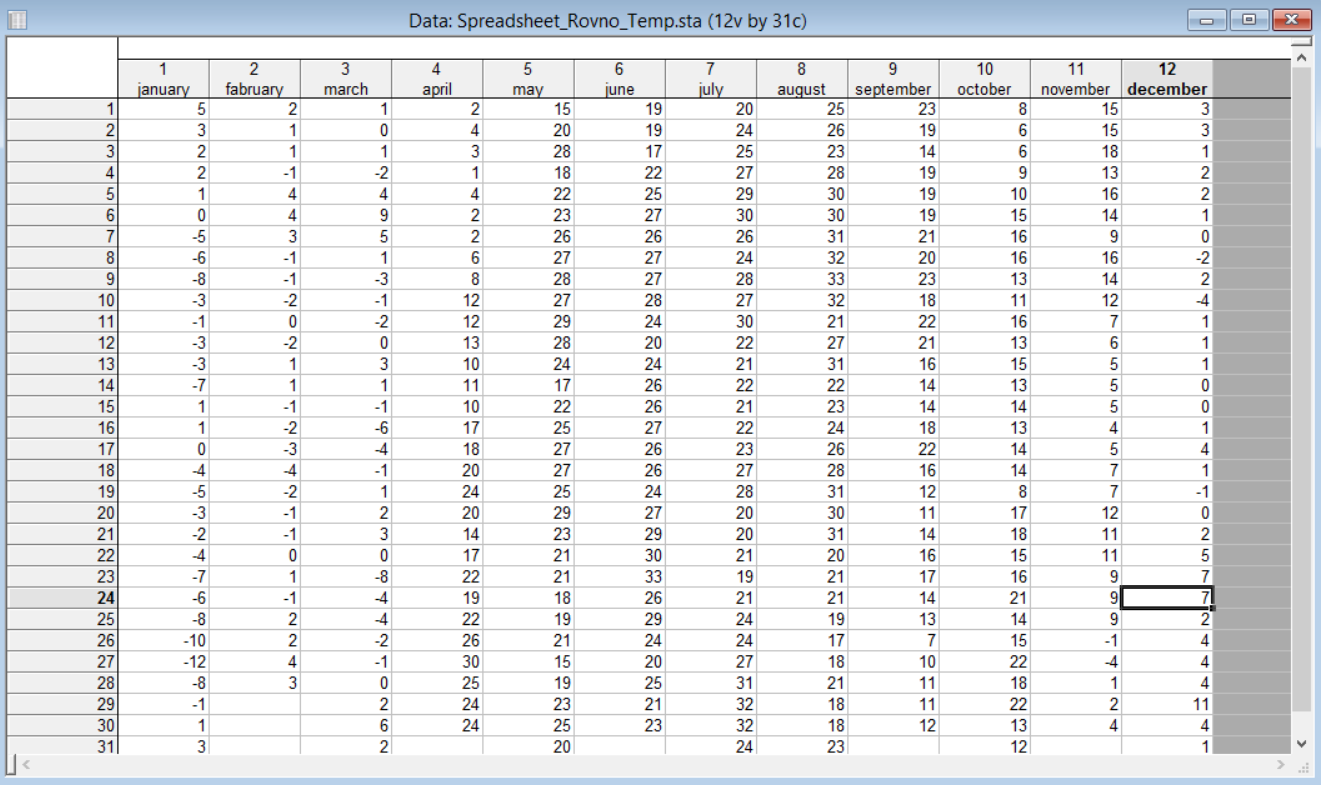


Рис. 2 - Початкові дані у Statistica

Тут змінним відповідають місяці року, а спостереженням - дні місяців. По своїй природі змінні неперервні, хоч на сайті "Ну і погода" й були представлені лише цілі значення температур. Метою нашого дослідження буде визначення можливих залежностей між змінними, побудова моделі для прогнозування температур по наявним даним.

# ОПИСОВИЙ АНАЛІЗ

Побудуємо графік температур квітня за 2013 рік:

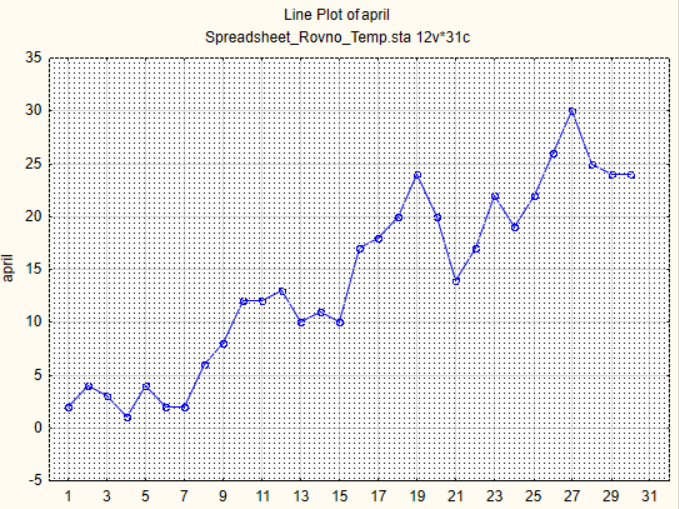


Рис. 3 - Графік температур квітня 2013

Бачимо, що температура протягом січня нерівно, але постійно зростає. Виконаємо те ж саме для січня:

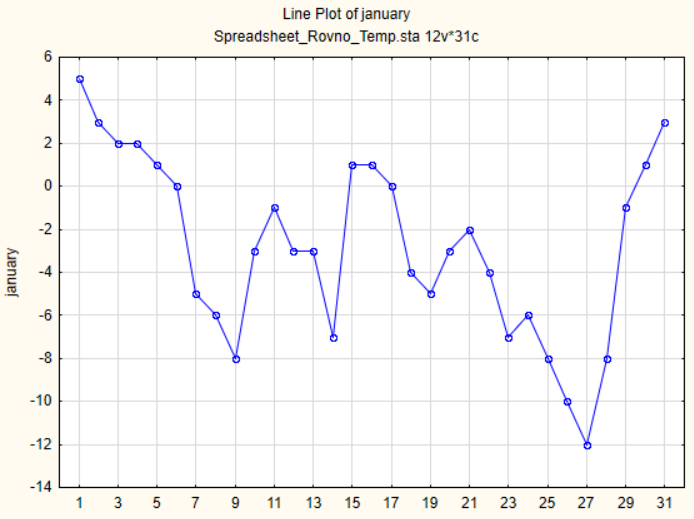


Рис. 4 - Графік температур січня 2013

Тепер побудуємо графіки всіх місяців кожної пори року. Результати представлені на рис. 5-8.

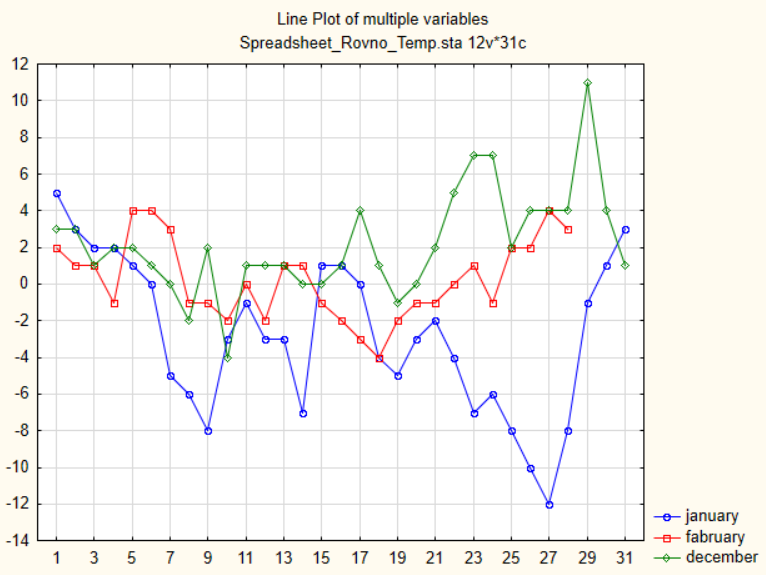


Рис. 5 - Графік температур зимових місяців

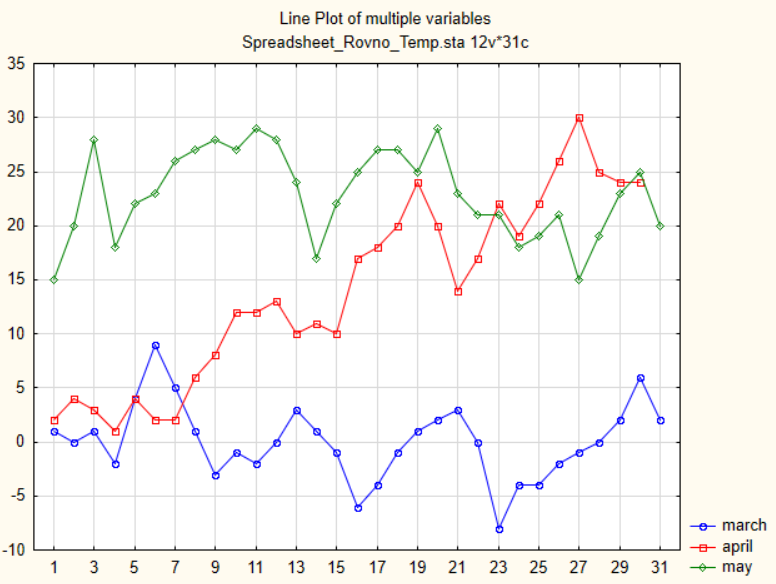


Рис. 6 - Графік температур весняних місяців

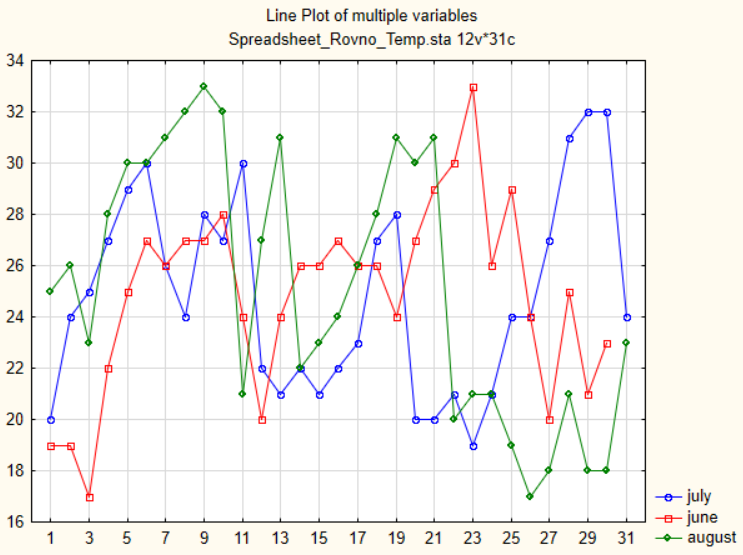


Рис. 7 - Графік температур літніх місяців

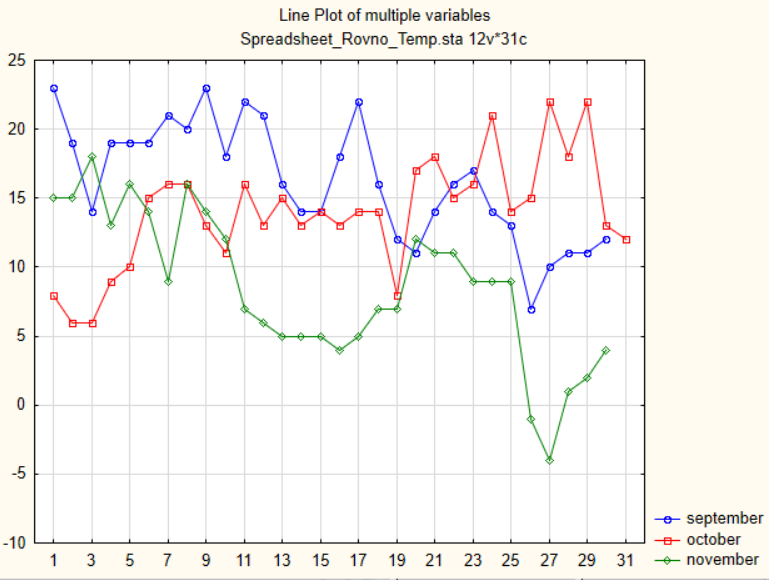


Рис. 8 - Графік температур осінніх місяців

Як бачимо, максимальне нарощення температури відбувається у квітні, а максимальний спад у листопаді. При цьому загалом літні й зимові температури більш нестабільні, ніж осінні та весняні.

Побудуємо графік середньомісячних температур по всьому періоду спостережень (рис. 9).

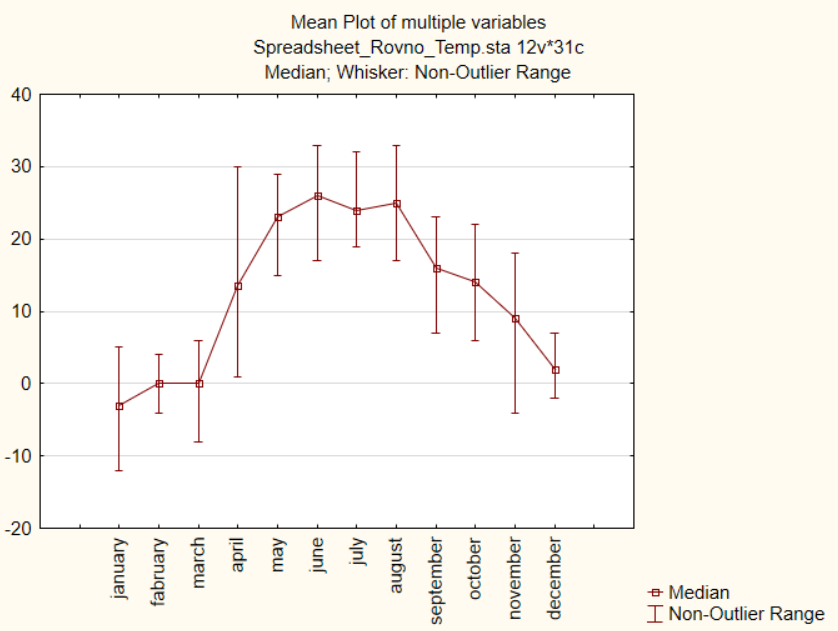


Рис. 9 - Графік середньомісячних температур за 2013 рік

Для виявлення взаємозалежностей у даних скористаємося асоціативними правилами. Для цього перетворимо початковий файл із даними (а точніше - створимо на його основі новий, в якому замість значень температур вкажемо одиниці для тих значень, які більше середнього для відповідного місяця, і нулі для тих, які менше). Щоб це зробити, треба спочатку визначити середні температури для кожного місяця (рис. 10).

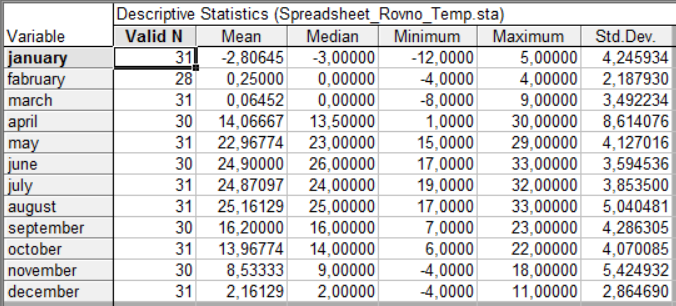


Рис. 10 - Описові статистики (в т.ч. середні температури місяців - mean)

Отже, новий файл із даними виглядає так:

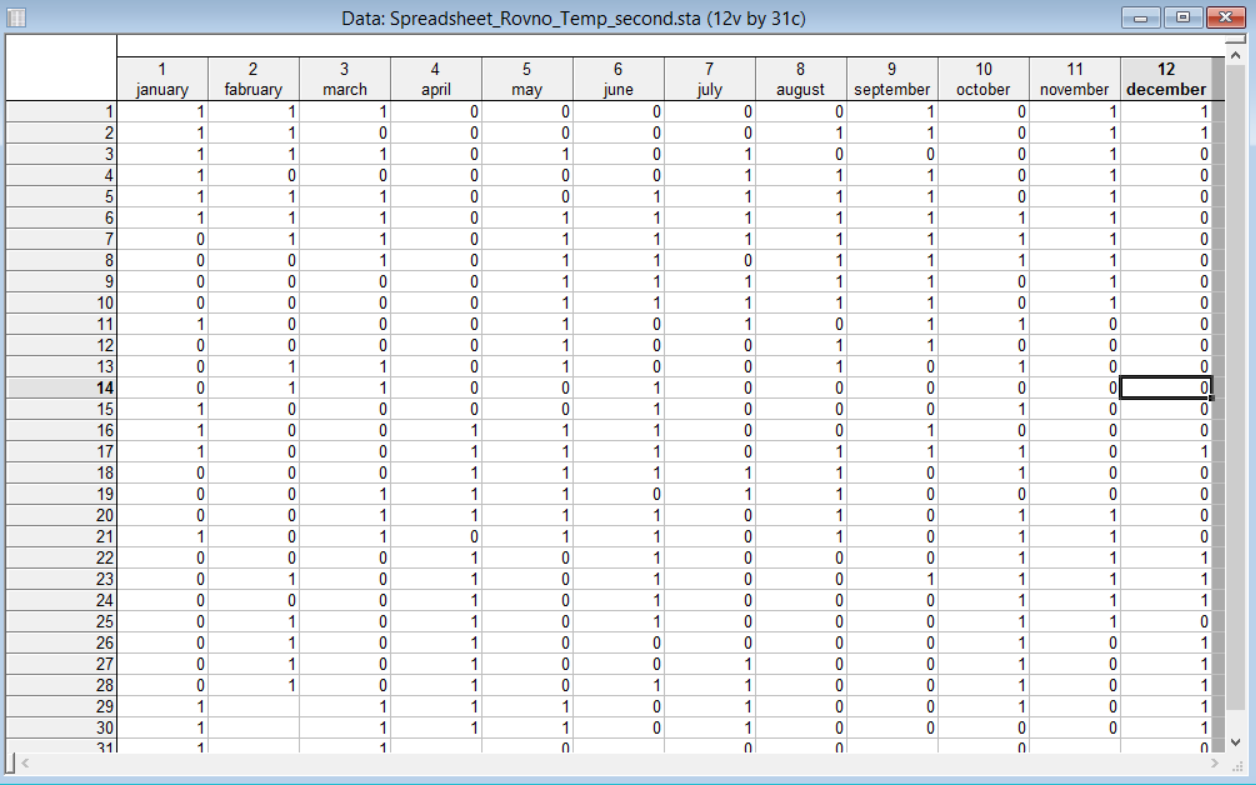


Рис. 11 - Новий файл із даними

Тепер знайдемо закономірності між даними за допомогою асоціативних правил (рис. 12).

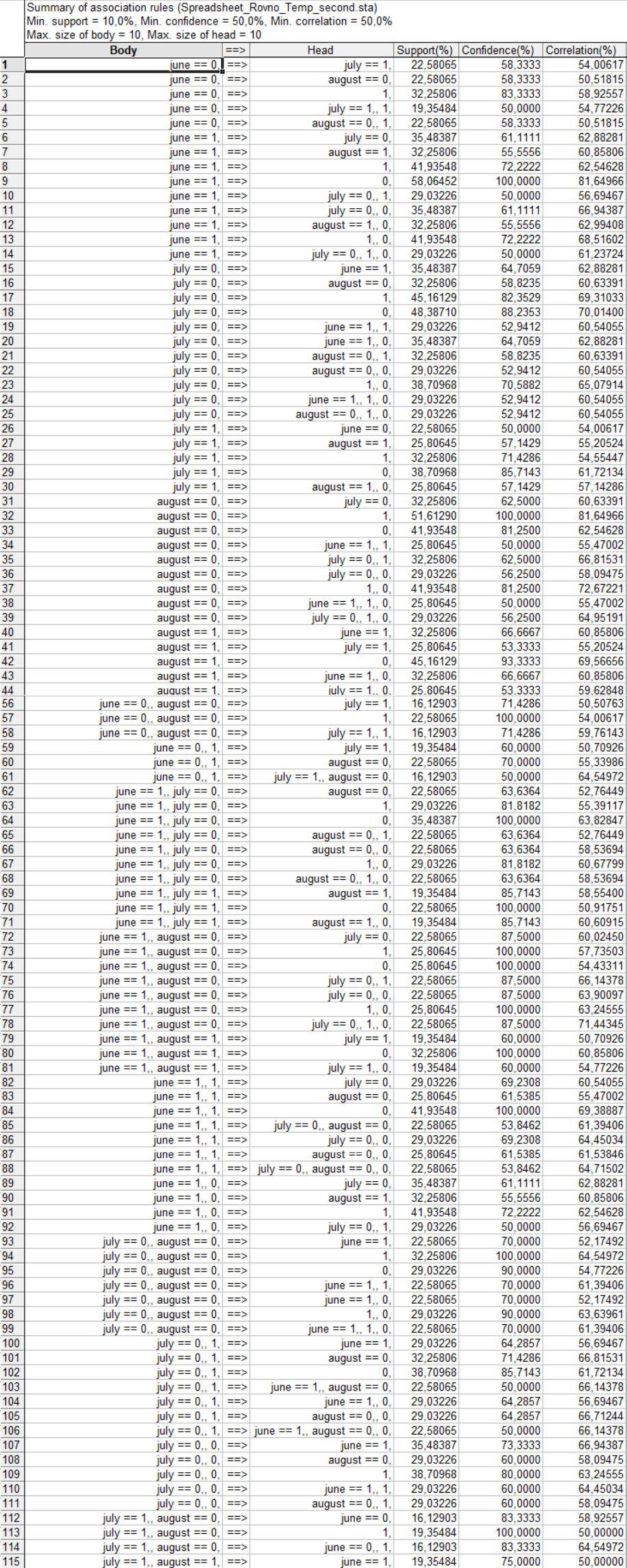


Рис. 12 - Результат застосування асоціативних правил

Бачимо, наприклад що якщо температура в липні у певні дні місяця більше середнього, то з великою ймовірністю температура у серпні для відповідних днів також буде більше середнього. Отримані результати - взаємозалежності між змінними - можна згодом використовувати при проведенні кластерного аналізу, побудові регресійних та прогнозових моделей.

# КЛАСТЕРНИЙ АНАЛІЗ

Виконаємо кластерний аналіз даних за допомогою методу k-середніх. Нам доведеться заздалегідь задати кількість кластерів (2), які ми хочемо виділити (див. рис. 13).

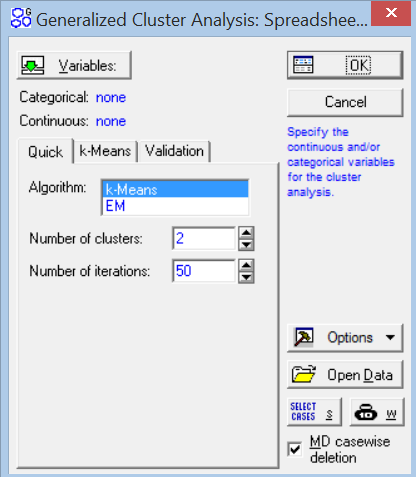


Рис. 13 - Налаштування кластерного аналізу

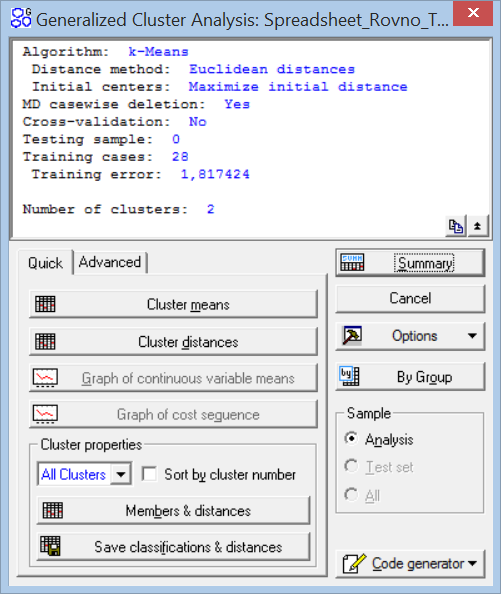


Рис. 14 - Результати аналізу

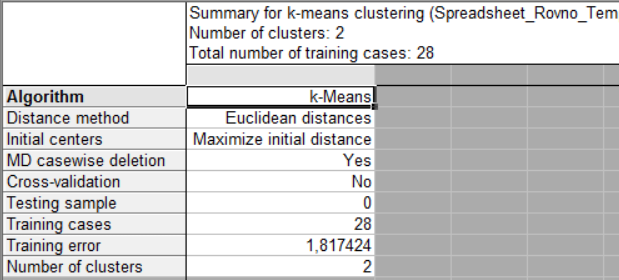


Рис. 15 - Підсумки по аналізу

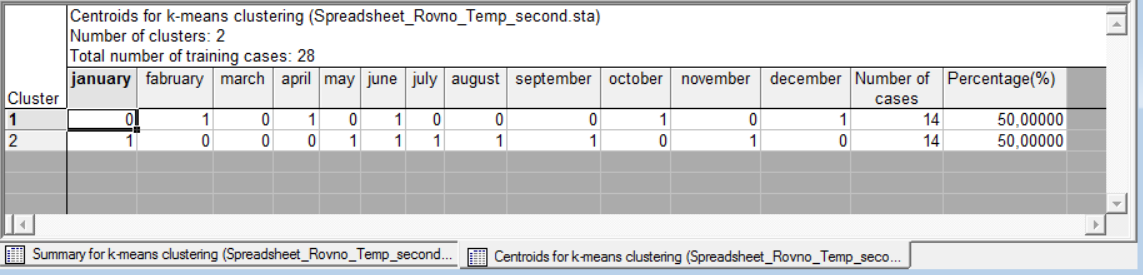


Рис. 16 - Середні значення змінних для двох знайдених кластерів

Бачимо, що, наприклад, для першого кластеру більш характерні температури серпня менше середнього (15,16), а для другого - більше середнього.

Тепер виконаємо аналогічний аналіз для початкових даних (де не нулі та одинички, а значення температур). Результати цього аналізу представлені на рис. 17-19.

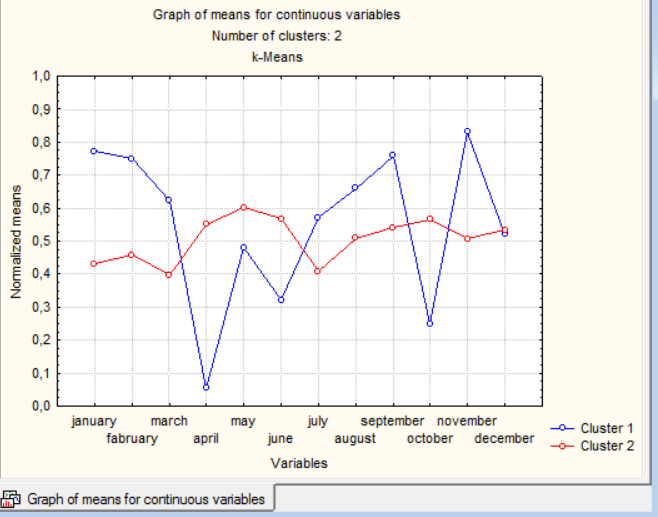


Рис. 17 - Графік нормалізованих середніх значень для двох кластерів

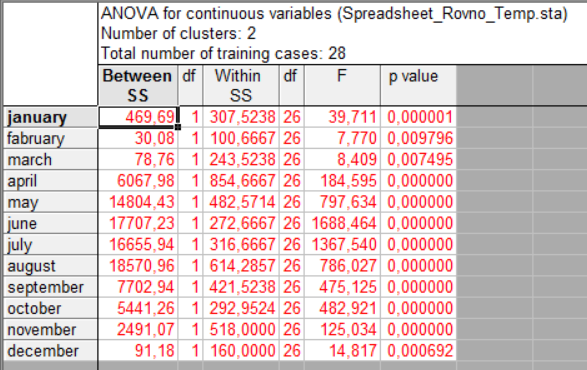


Рис. 18 - Результати дисперсійного аналізу

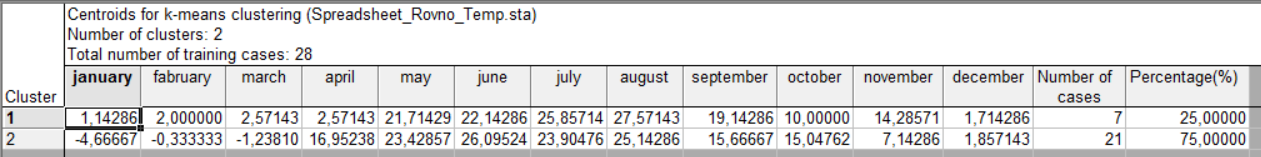


Рис. 19 - Середні значення температур місяців для двох кластерів

Таким чином, всі місяці мають значний вплив на розділення постережень на кластери; перший кластер містить 7 спостережень (днів), другий - 21; середнє значення температури жовтня для першого кластера становить 14,29, а для другого кластера - 7,14.

# ПРОГНОЗУВАННЯ ТЕМПЕРАТУРИ НА 20 ЛЮТОГО 2014 РОКУ

Спочатку спрогнозуємо температуру на всі дні лютого 2013 року. Для цього скористаємося методом Support Vector Machines (методом опорних векторів) (див. рис. 20).

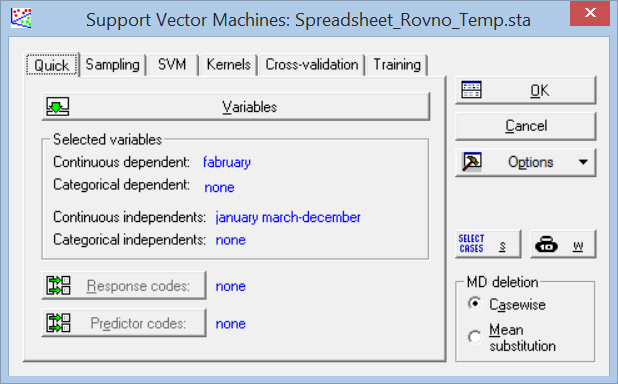


Рис. 20 - Метод Support Vector Machines

Перші 22 спостереження візьмемо для навчання вибірки, інші - для перевірки прогнозів (рис. 21).

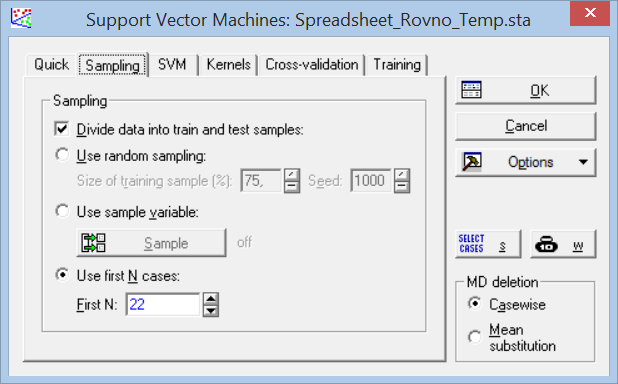


Рис. 21 - Розділення даних на тестові та тренувальні

Результати прогнозування представлені на рис. 22-24.

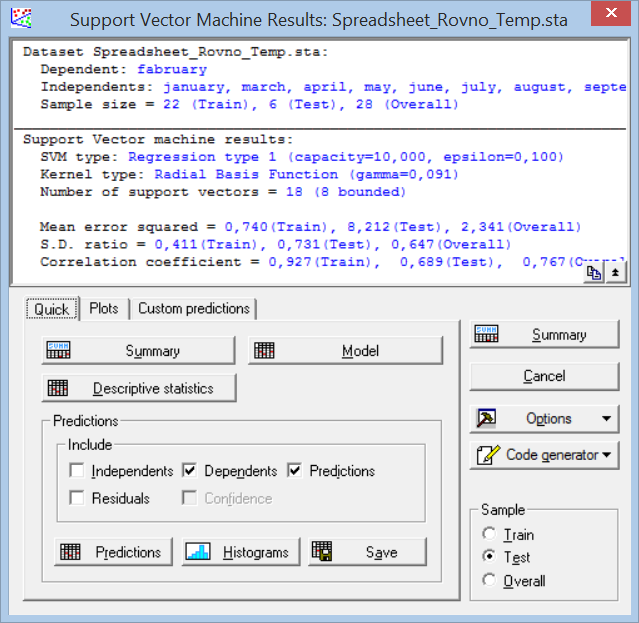


Рис. 22 - Результати прогнозування

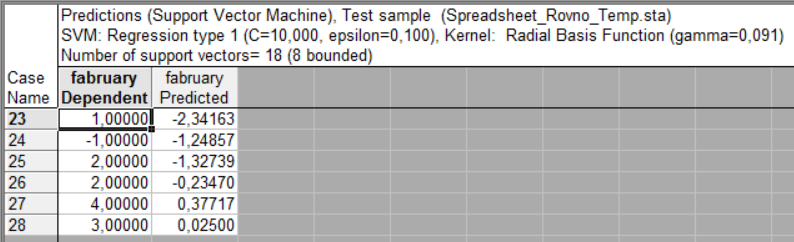


Рис. 23 - Результати порівняно з відомими значеннями для тестових спостережень

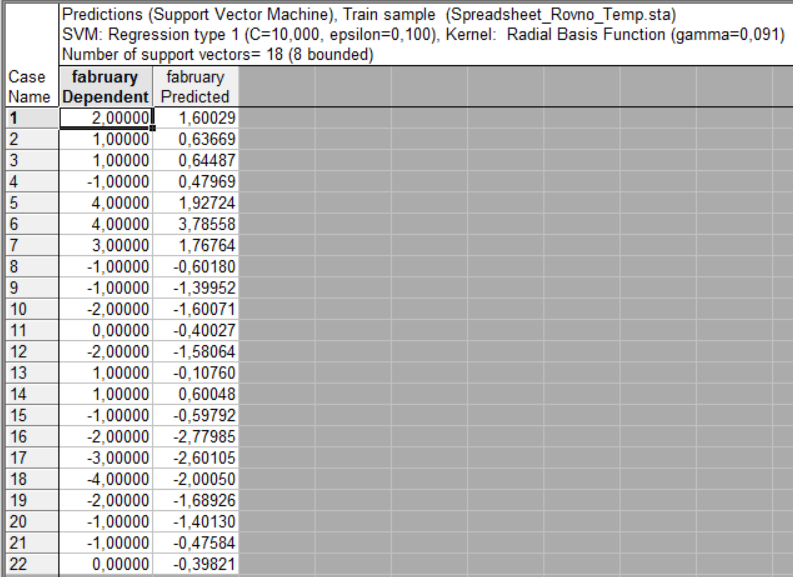


Рис. 24 - Результати порівняно з відомими значеннями для тренувальних спостережень