Зміст

[1. Вступ 2](#_Toc471930581)

[2. Постановка задачі 3](#_Toc471930582)

[3. Аналіз даних 5](#_Toc471930583)

[4. Прогнозування координат точок на площині 10](#_Toc471930584)

[1. Метод k найближчих сусідів 10](#_Toc471930585)

[5. Оцінка точності алгоритму 11](#_Toc471930586)

[6. Висновок 14](#_Toc471930587)

# Вступ

\*гарна вступна промова\*

**Мета роботи**. Дослідити та порівняти методи прогнозування послідовного вибору точок на двовимірної площині.

**Ціль роботи.** Створити найефективніший алгоритм для приблизного передбачення послідовного вибору точок на площині користувачем.

**Актуальність**. Базуючись на задачі прогнозування поведінки людей на площині, здається можливим перейти до задачі прогнозування переміщення тих чи інших людей по встановленим секторам міста, що може бути корисним для правоохоронних органів при пошуку злочинців або загублених людей.

# Постановка задачі

**Збір даних.** Дані були зібрані окремо з кожного респондента за допомогою комп’ютерного додатку. Користувач мав обрати на площині 15 випадкових точок, які записувалися у цілісний блок даних для кожного респондента. Усі точки мали цілочисленні координати у межах від 0 до 600. У експерименті приймали участь учні ЛІТу віком від 14 до 16 років.

Деякі дані, наприклад, такі, у яких дисперсія за певним показником не перевищувала 100, були відсіяні. Також були відфільтровані дані з очевидною структурою (симетричні, ті, що складаються у якусь фігуру і т.д.) і дані, які не відповідали умовам задачі (наприклад, такі, у яких координати точок виходили за межі досліджуваного діапазону).

**Перехід до задачі кидання двох точок на відрізок.** Нехай стохастичний експеримент полягає у киданні навмання точки в множину з . Надалі в слова «точку навмання кидають у множину » ми вкладатимемо такий зміст: кинута точка може потрапити в будь-яку підмножину множини , і ймовірність того, що точка опиниться в , пропорційна мірі Лебега множини . Формально це означає, що як математичну модель стохастичного експерименту, що полягає у киданні навмання точки у множину , розглядатимемо , де  **–** клас борелевих підмножин , – імовірність на класі , яка для кожного означається рівністю:

де – міра Лебега на .

Помітимо, що упорядкованій парі точок на відрізку з коодинатами та відповідає одна точка у квадраті з координатами і навпаки. Тому вибір випадкової точки у цьому квадраті рівносильний вибору двох випадкових точок на відрізку . Таким чином можна перейти від задачі вибору точок на площині до задачі вибору точок на відрізку.

**Математична модель.** Нехай – множина респондентів (користувачів), а – множина точок, які були обрані респондентом. Кожний респондент має можливість обрати задану кількість точок з цілочисленими координатами на квадратній площині.

Введемо матрицю

де елемент – координати точки під номером респондента . Загалом, частина точок відома (задана), а частина – невідома.

Отже, задача заключається у оцінці потенціальних значень невідомих елементів за вже відомими (заданими) елементами.

# Аналіз даних

Дискретна випадкова величина описується своїм розподілом

У загальній ситуації випадкова величина описується так званою функцією розподілу, за якою завжди можна обчислити імовірність того, що потрапить до обраного проміжку .

Ця функція визначена так:

Якщо функцію розподілу випадкової величини можна подати у вигляді

то кажуть, що випадкова величина має абсолютно неперервний розподіл (абсолютно неперервна), а функцію називають щільністю розподілу випадкової величини .

Випадкова величин має *нормальний розподіл із параметрами* , якщо її щільність розподілу

**Зображення даних.** На рисунку 1 зображено усі дані в сукупності – кожному кольору відповідає певний респондент. Загалом, очевидних закономірностей розташування точок немає.

На другому графіку k позначає номер точки для кожного респондента. Можна помітити, що зі збільшенням k збільшується хаотичність у виборі точок, а на початку вони зосереджені у одній області. Можна зробити висновок: зі збільшенням кількості точок дисперсія збільшується.

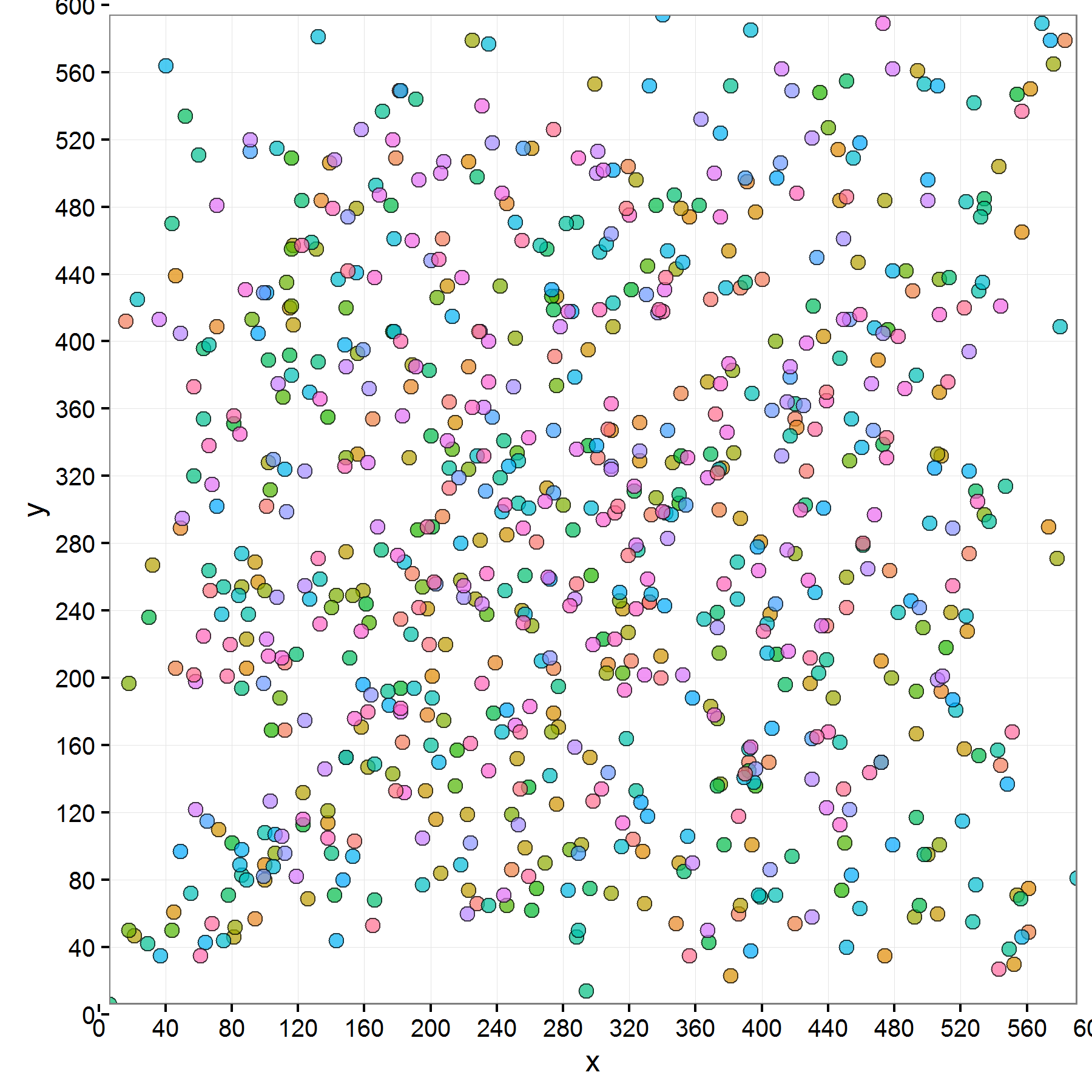


Рис. 1. Загальне зображення даних.

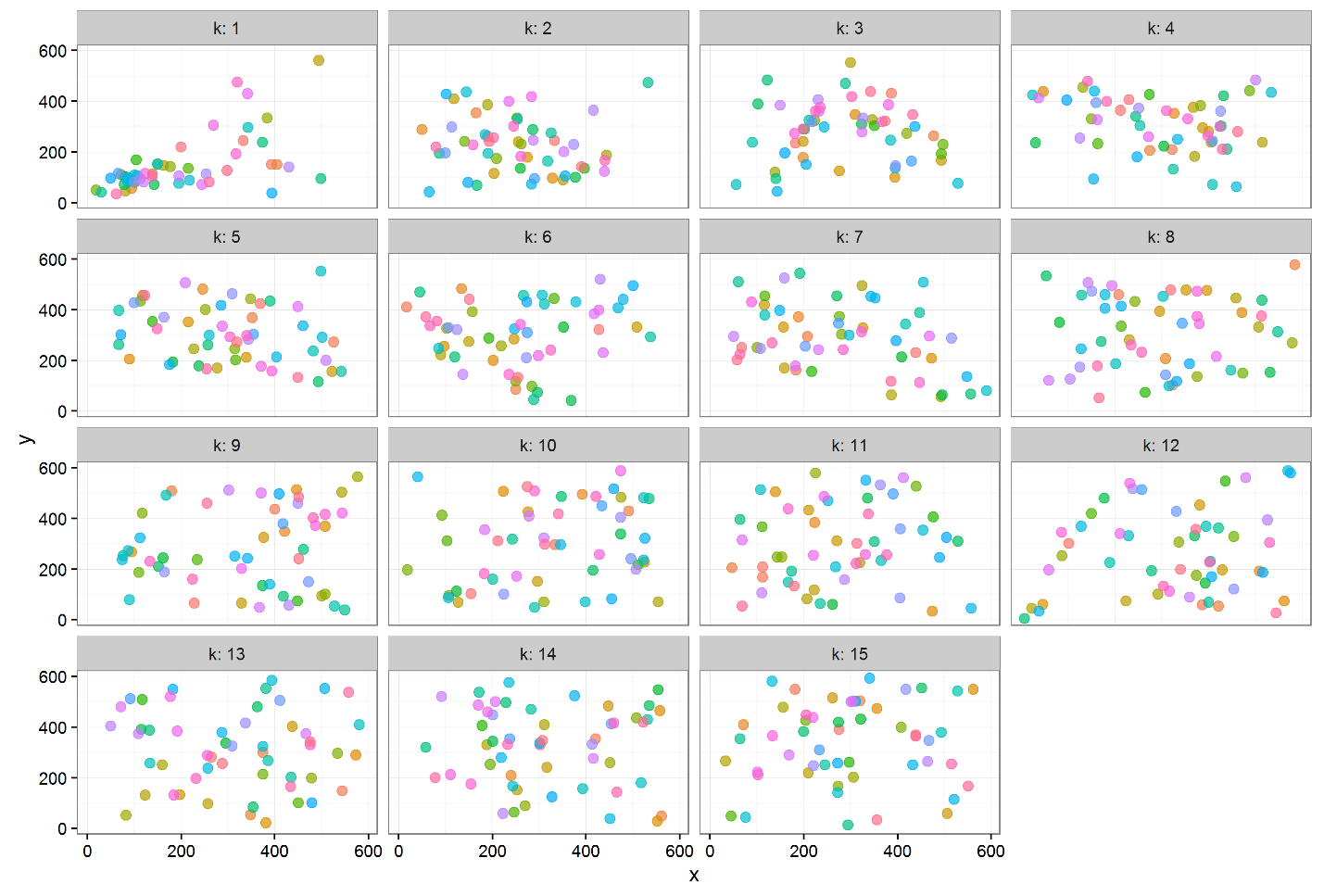


Рис. 2. Зображення даних по крокам.



Рис. 3. Зображення даних по респондентам.

**Перевірка даних на нормальність.** Перевірятимемо дані на нормальність окремо за кожною координатою.

Аналіз щільності, зображений на рисунках 4 та 5, надає підстави висунути припущення щодо нормального розподілу зібраних даних. Пояснити таке припущення можливо завдяки тому, що здебільшого функція щільності набуває найбільшої маси в середині відрізку [0; 600], тобто розподіл має коефіцієнт асиметрії близький до нуля, та середнє значення близьке до моди та медіани.

Для додаткового аналізу наведено рисунки 6 та 7, що містять так звані графіки «квантиль-квантиль». Ці графіки надають ще більше впевненості в близкості зібраних даних до нормального розподілу.

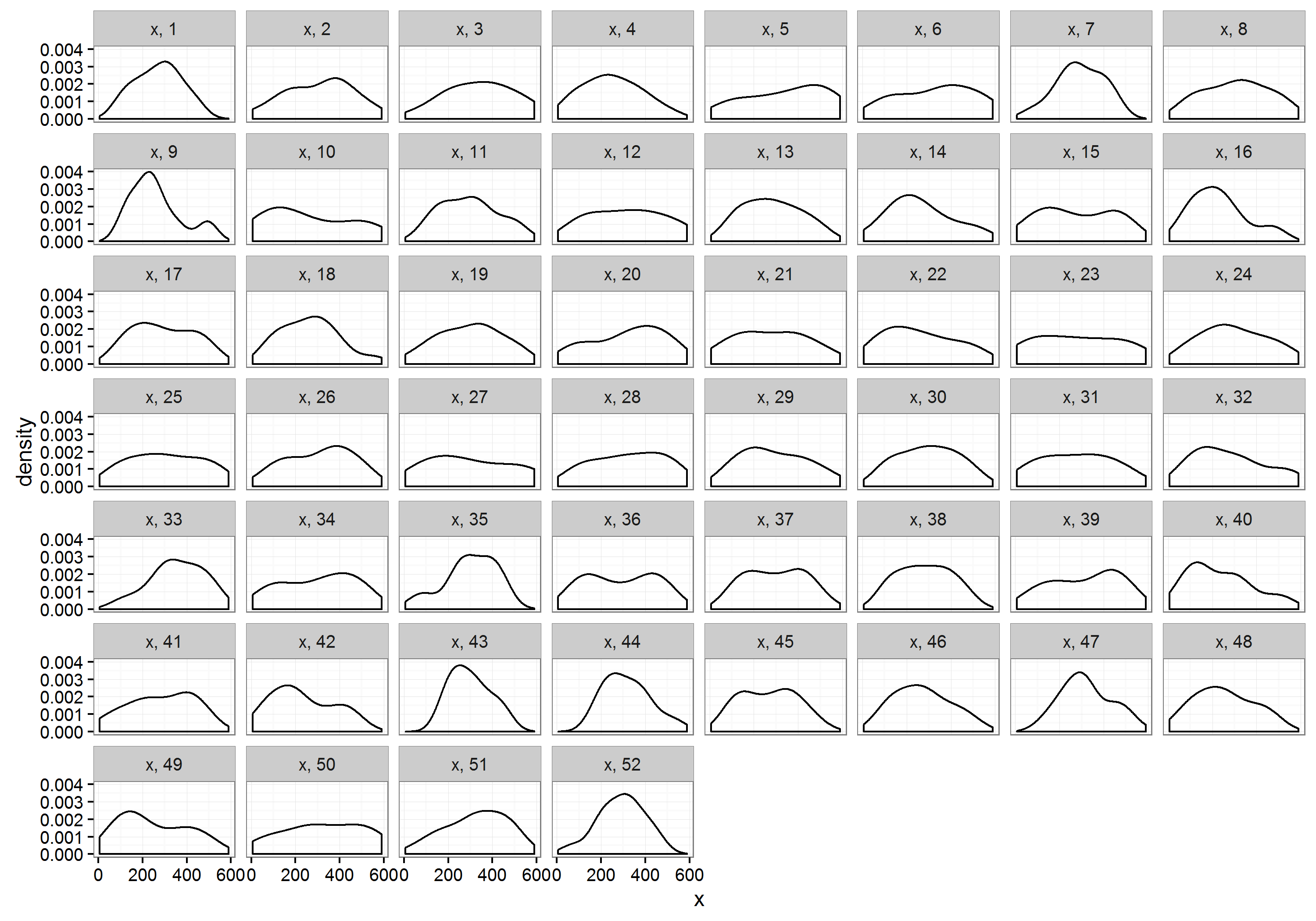


Рис. 4. Графіки щільності координат x за респондентами.

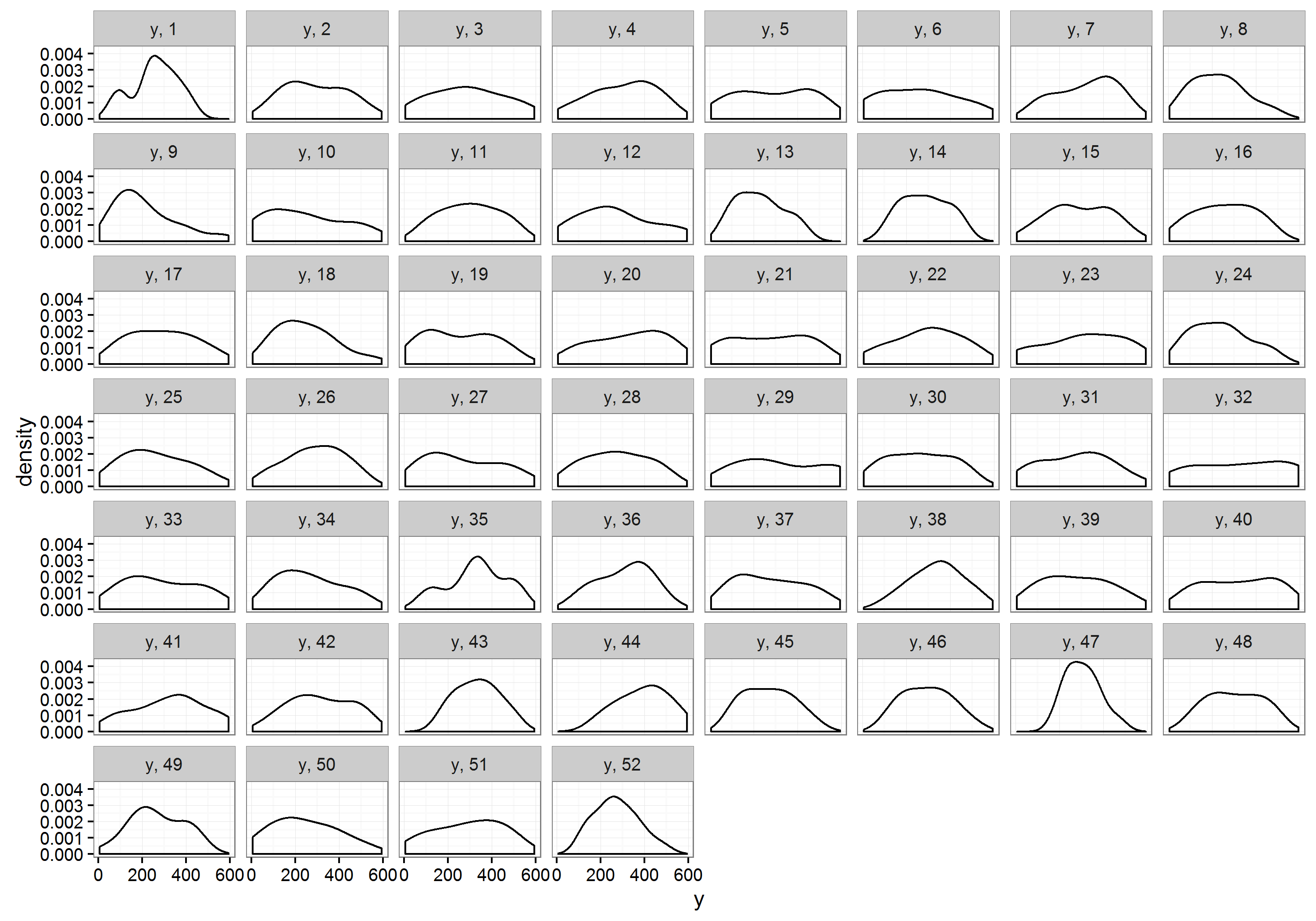


Рис. 5. Графік щільності координат y за респондентами.

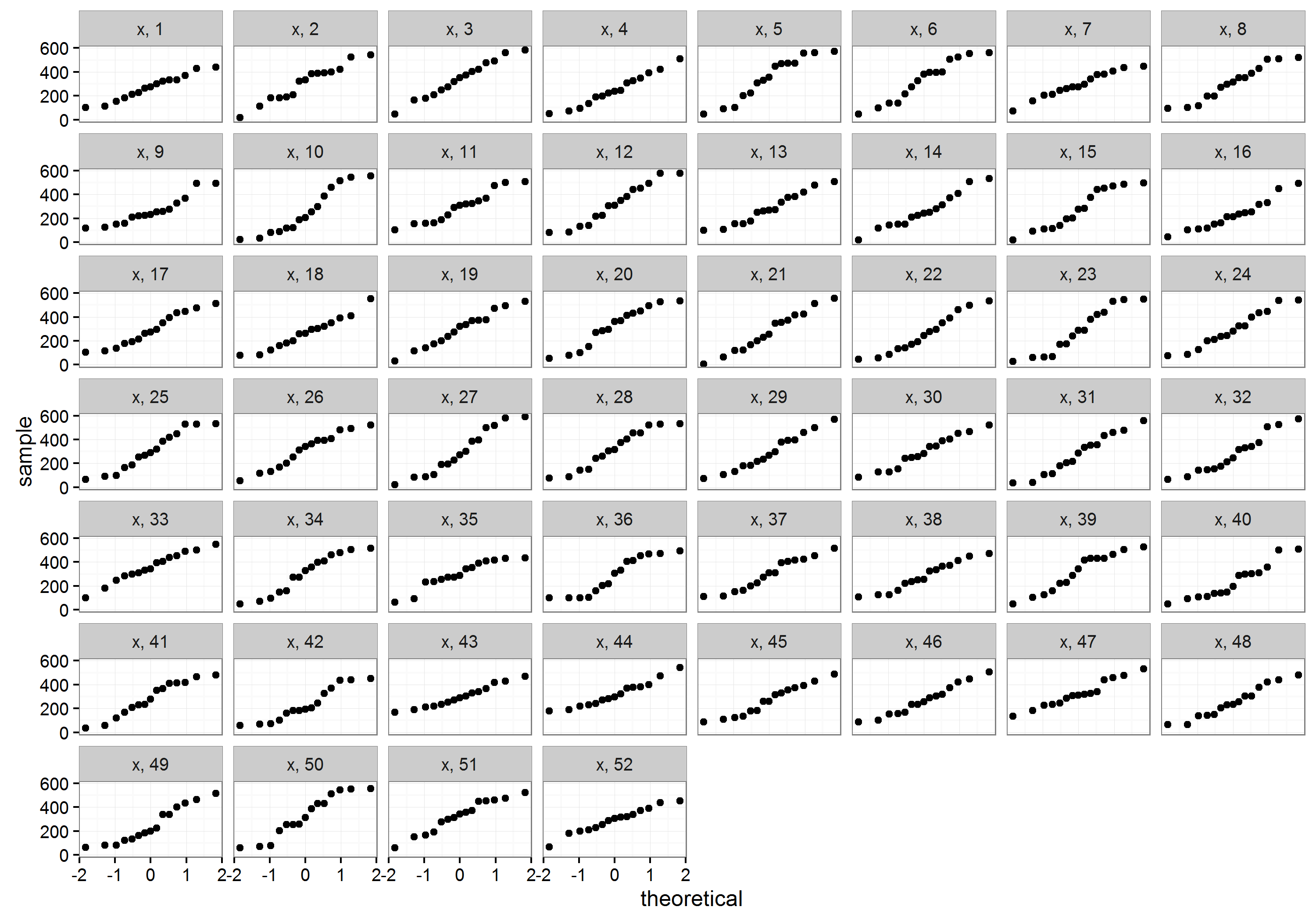


Рис 6. Графік квантиль-квантиль за координатою x респондентів.

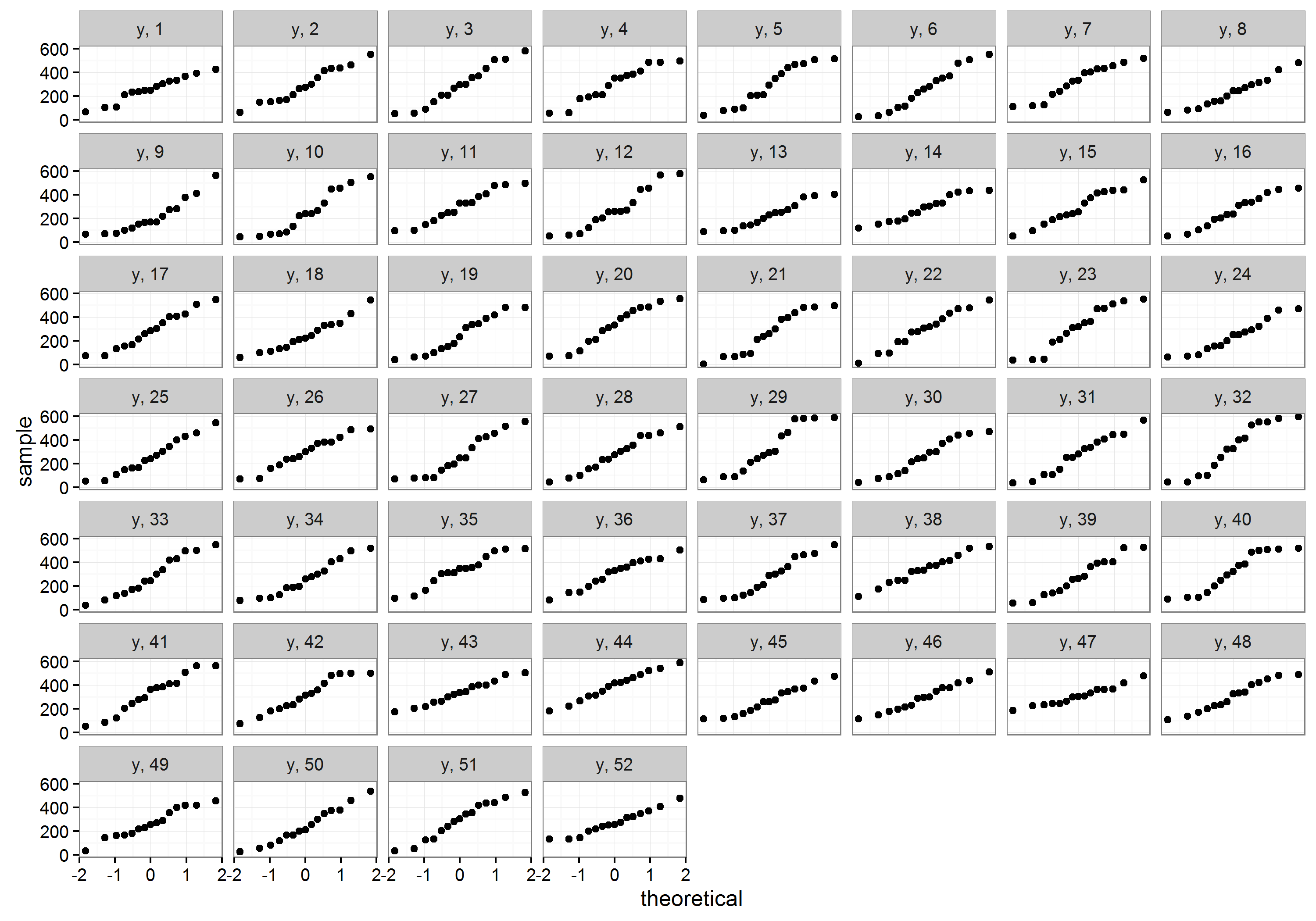


Рис 7. Графік квантиль-квантиль за координатою y респондентів.

# Прогнозування координат точок на площині

Колаборативна фільтрація – метод, за якого невідомі елементи множини оцінюються лише за відомими елементами тієї ж множини без викроистання додаткової інфромації.

Ідея колаборативної фільрації заключається у тому, що «схожі» люди діють подібним чином.

Найбільш використовуваний метод для данного способу – к найближчих сусідів (*K Nearest Neighbors*).

## Метод k найближчих сусідів

Мірою подібності респондентів може слугувати будь-яка метрика на просторі m-мірних векторів, наприклад, індукована евклідова метрика, коефіцієнт кореляції Пірсона (показник лінійної залежності між центрованими векторами), косинуса подібності (показник лінійної залежності між векторами) і т.д. Розрізняють два варіанти реалізації методу к найближчих сусідів: орієнтований на користувачів та орієнтований на об’єкти.

Виходячи з формулювання досліджуваної задачі, розглядатимемо реалізацію, орієнтовану на користувачів.

Ідея методу: знаючи m-вимірні вектори послідовного вибору точок на площині для кожного респондента, можна встановити між ними міру подібності та представити невідоме значення точки кожного користувача через лінійну комбінацію (зважену суму) його сусідів.

Коефіцієнт кореляції Пірсона між користувачами розраховується з урахуванням спроможної та незміщеної оцінки дисперсії:

Коефіцієнт кореляції Спірмена між двома користувачами розраховується за формулою:

де

Виходячи зі специфіки задачі, оцінювати координати невідомих точок будемо окремо для координати x та для координати y.

З коефіцієнтів подібності складається матриця подібності розміром , де – кількість користувачів. Кожен коефіцієнт матриці – кількісна міра подібності між користувачами та .

Будемо обирати для кожного користувача найбільших коефіцієнтів з матриці , які відповідають найбільш схожим на нього користувачам. Коефіцієнт обирається довільно.

Невідомі значення елементів де – значення точки для найближчих сусідів користувача .

# Оцінка точності алгоритму

Найпоширенішим підходом для оцінювання точності алгоритму є розрахунок середньоквадратичної похибки алгоритму.

, де – множина відомих точок користувачів, а – множина наближених оцінок відповідно відомих точок.

Використовуючи коефіцієнти Пірсона і Спірмена, були отримані наступні значення середньоквадратичної похибки:

|  |  |  |
| --- | --- | --- |
| Кількість сусідів | RMSE(КК Пірсона) | RMSE(КК Спірмена) |
| 1 | 70.243 | 76.819 |
| 2 | 77.931 | 84.020 |
| 3 | 71.962 | 77.724 |
| 4 | 69.984 | 75.145 |
| 5 | 68.851 | 74.476 |
| 10 | 67.262 | 73.096 |
| 15 | 67.423 | 73.047 |
| 20 | 67.882 | 73.282 |
| 25 | 68.176 | 73.401 |
| 30 | 68.434 | 73.656 |
| 35 | 68.785 | 73.829 |
| 40 | 69.064 | 74.145 |
| 45 | 69.378 | 74.353 |
| 50 | 69.654 | 74.603 |

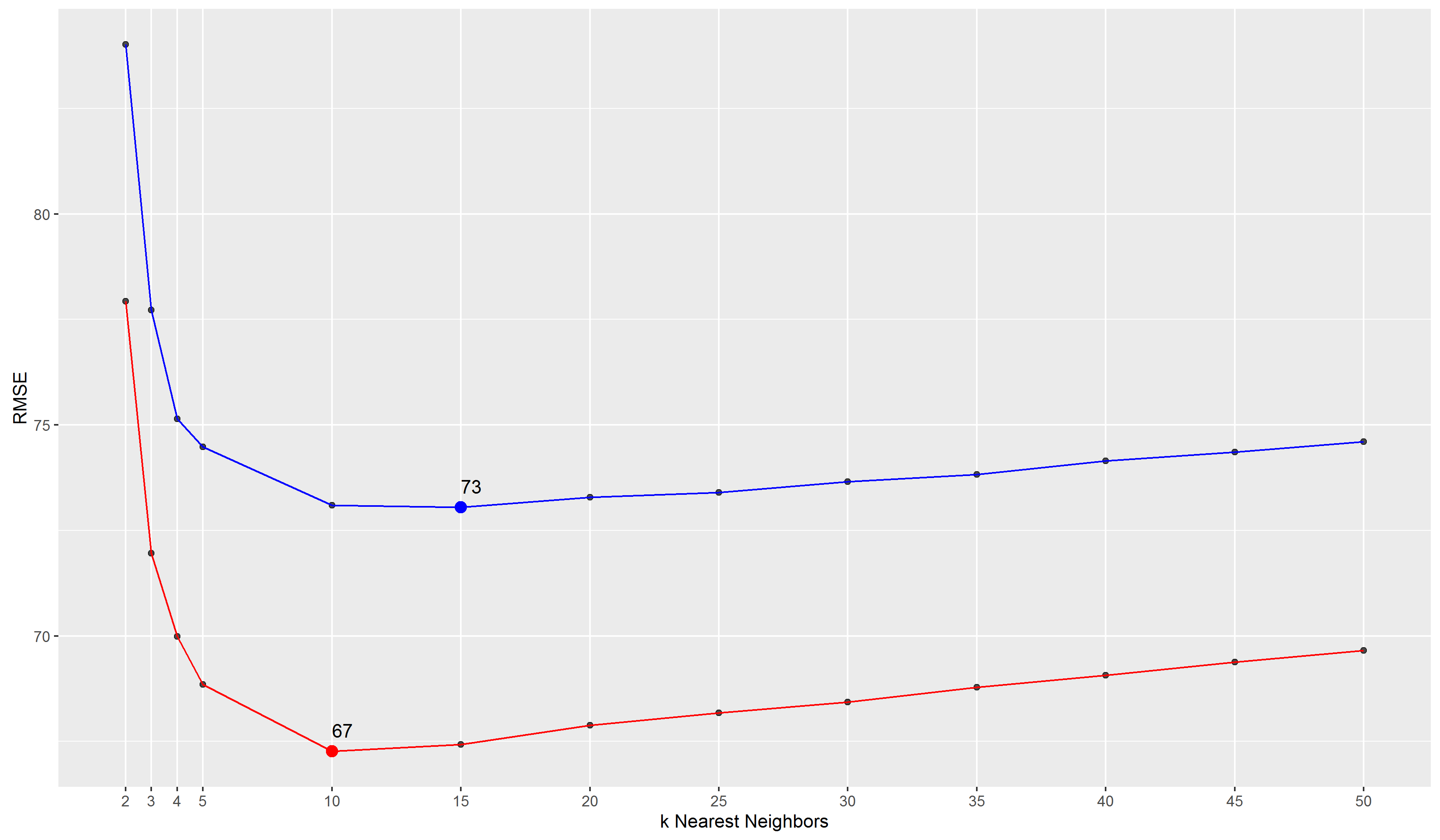


Рис 7. Залежність RMSE від k для КК Спірмена і Пірсона.

Як можна побачити з графіку, коефіцієнт кореляції Пірсона (червона лінія) дає найкращі результати при k=10.

Для порівняння двух реалізацій методу порівняємо значення середньоквадратичних помилок цих реалізацій. Для цього використаємо критерій Стьюдента про рівність середніх при невідомих дисперсіях. Право на використання критерія Стьюдента нам надає центальна гранична теорема, котра говорить, що середнє значення випадковохої величини, отриманої з будь-якого розподілу, має нормальний розподіл.

Нехай та – незалежні виборки відповідно з нормльного розподілу і , – верхня межа розподілу Стьюдента з (n+m-2) степенями свободи. Якщо гіпотезу відхиляти при

і не відхиляти у протилежному випадку, то з імовірністю гіпотеза не приймається, коли вона правильна. Тут , , , і

Розв’язуючи задачу, отримуємо результат .

# Висновок