

Обробка та аналіз даних про ґрунт з використанням PANDAS та MATPLOTLIB

Іван Гребенюк,

здобувач вищої освіти спеціальності 122 Комп'ютерні науки

Миколаївський національний аграрний університет

м. Миколаїв, Україна

Анотація: дослідження присвячено розгляду методів обробки та аналізу ґрунтових даних за допомогою бібліотек мови програмування Python Pandas і Matplotlib. Розглядаються методи структурування, візуалізації та виявлення закономірностей у великих наборах даних.

Ключові слова: ґрунтові дані, Pandas, Matplotlib, аналіз, візуалізація, машинне навчання, сільське господарство.

Ґрунт є основним ресурсом, який визначає продуктивність сільського господарства, лісового господарства та загальну екологічну стабільність. Важливість ґрунту в цих сферах зумовлює необхідність його ретельного аналізу, що включає оцінку якості, кількості та придатності для різних видів використання. Ґрунт не лише забезпечує умови для росту рослин, але й виконує ключові екологічні функції, такі як зберігання вуглецю, очищення води та підтримка біорізноманіття.

Сучасні програмні інструменти, такі як Pandas та Matplotlib, суттєво полегшують процес обробки та аналізу великих обсягів даних, що стосуються ґрунтових досліджень. Pandas, потужна бібліотека для маніпуляції з даними, дозволяє ефективно обробляти, фільтрувати та аналізувати інформацію, що стосується характеристик ґрунту. За допомогою Pandas можна виконувати різноманітні статистичні аналізи, що сприяють кращому розумінню властивостей ґрунту та його потенціалу для різних видів використання.

У свою чергу, Matplotlib надає можливість візуалізації даних, що дозволяє наочно представити результати аналізу. Візуалізація даних є ключовим етапом у дослідженнях, оскільки вона допомагає виявити тренди, закономірності та аномалії, що можуть залишитися непоміченими при простому перегляді числових показників.

Pandas є однією з найпотужніших бібліотек Python для обробки та аналізу даних. Вона забезпечує високорівневі структури даних та інструменти для роботи з таблицями та часовими рядами. Основними структурами даних у Pandas є DataFrame та Series, які дозволяють зручно маніпулювати табличними даними.

Для початку роботи з даними про ґрунт необхідно завантажити їх у форматі, зручному для аналізу. Pandas підтримує різні формати файлів, такі як CSV, Excel та SQL. Наприклад, для завантаження даних з CSV-файлу використовується функція `read_csv`:

```
import pandas as pd
```

```
data = pd.read_csv('soil_data.csv')
```

Після завантаження даних важливо провести їх попередню обробку, яка включає обробку пропущених значень – видалення або заповнення відсутніх даних, перетворення типів даних – забезпечення правильних типів для кожного стовпця, нормалізацію та масштабування – приведення даних до єдиного масштабу для подальшого аналізу.

Важливим кроком у аналізі даних про ґрунт є виявлення взаємозв’язків між різними параметрами, такими як рН, вміст органічної речовини, макро- та мікроелементи. Pandas надає функцію `corr()`, яка обчислює кореляційну матрицю для числових стовпців: `correlation_matrix = data.corr()`

Такий аналіз допомагає визначити, які параметри мають сильний взаємозв’язок, що може бути корисним для прогнозування та прийняття рішень.

Matplotlib є основною бібліотекою для візуалізації даних у Python. Вона дозволяє створювати різноманітні графіки та діаграми для наочного представлення інформації. Для розуміння розподілу окремих параметрів ґрунту, таких як рН або вміст азоту, корисно використовувати гістограми, наприклад:

```
import matplotlib.pyplot as plt
plt.hist(data['pH'], bins=20, edgecolor='black')
plt.title('Розподіл рН ґрунту')
plt.xlabel('pH')
plt.ylabel('Частота')
plt.show()
```

Для візуалізації взаємозв’язків між двома параметрами зручно використовувати діаграми розсіювання:

```
plt.scatter(data['pH'], data['organic_matter'])
plt.title('Залежність між рН та вмістом органічної речовини')
plt.xlabel('pH')
plt.ylabel('Вміст органічної речовини (%)')
plt.show()
```

Такий підхід дозволяє виявити лінійні або нелінійні залежності між параметрами.

Окрім базового аналізу, бібліотеки Pandas та Matplotlib можуть бути використані у поєднанні з бібліотекою `scikit-learn` для побудови моделей машинного навчання, що прогнозують певні властивості ґрунту на основі наявних даних.

Перед побудовою моделі необхідно підготувати дані: зробити вибір ознак та цільової змінної, тобто визначитись, які параметри будуть використовуватися як вхідні дані, а які – як вихідні, а також розбити дані на тренувальний та тестовий набори для оцінки точності моделі. Наприклад, для прогнозування вмісту органічної речовини на основі інших параметрів можна використовувати лінійну регресію:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
```

```

# Вибір ознак та цільової змінної
X = data[['pH', 'nitrogen', 'phosphorus', 'potassium']]
y = data['organic_matter']
# Розбиття на тренувальний та тестовий набори
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
# Створення та тренування моделі
model = LinearRegression()
model.fit(X_train, y_train)
# Прогнозування та оцінка моделі
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
print(f'Середньоквадратична помилка: {mse}')

```

Такий підхід дозволяє створювати прогностичні моделі, які допомагають аграріям приймати обґрунтовані рішення щодо удобрення та обробки ґрунту.

Результати великої кількості досліджень доводять ефективність Pandas і Matplotlib у процесах обробки та аналізу ґрунтових даних. Вони забезпечують швидке опрацювання, статистичну оцінку та якісну візуалізацію, що сприяє обґрунтованим рішенням у сфері управління земельними ресурсами. Використання сучасних комп'ютерних методів аналізу ґрунту дозволяє покращити агротехнології та екологічний моніторинг. Подальші дослідження можуть охоплювати складніші алгоритми машинного навчання для більш точних прогнозів.

Список використаних джерел:

1. Saini, M., & Kumar, R. (2019). Exploring Data Visualization and Analysis with Matplotlib. In International Journal of Psychosocial Rehabilitation (pp. 2189–2194). Ninety Nine Publication. <https://doi.org/10.61841/v23i4/400326>
2. Diana Julie M, D. (2023). Exploring the Paradigm Shift: Harnessing Data Analytics for Real - World Applications. In International Journal of Science and Research (IJSR) (Vol. 12, Issue 6, pp. 1467–1480). International Journal of Science and Research. <https://doi.org/10.21275/sr23611121501>

Abstract: *The study is devoted to the methods of processing and analysing soil data using the Python programming language libraries Pandas and Matplotlib. The methods of structuring, visualisation and detection of patterns in large data sets are considered.*

Keywords: *soil data, Pandas, Matplotlib, analysis, visualization, machine learning, agriculture.*

Науковий керівник:

Пархоменко О. Ю.,

*к. ф.-м. н., доцент кафедри економічної кібернетики,
комп'ютерних наук та інформаційних технологій,
Миколаївський національний аграрний університет*