**Задача оптимального выбора комбинации экзогенных переменных для моделей SARIMAX с использованием Lasso-регрессии**

***А.И. Балюк***

*Студент
Московский государственный университет имени М.В.Ломоносова,
физический факультет, Москва, Россия
E-mail: baliuk.ai20@physics.msu.ru*

В данной работе исследуется использование моделей SARIMAX для прогнозирования временных рядов с экзогенными переменными и оптимизация их выбора с применением Lasso-регрессии. Проведенный анализ данных Тверской области показал, что отбор значимых экзогенных переменных с использованием Lasso-регрессии позволяет минимизировать ошибку прогноза и предотвратить переобучение модели. Полученные результаты подтверждают, что правильный выбор экзогенных переменных существенно улучшает качество предсказательной модели.

Прогнозирование временных рядов и заполнение пропусков данных являются важными задачами в прикладной науке. В работе исследуется один из основных инструментов прогнозирования - статистическая модель SARIMAX, являющаяся расширением модели ARMA - а именно: включающая интегрирующую, сезонную компоненты и экзогенные переменные. Однако избыточное число экзогенных переменных может привести к переобучению, и, как следствие, возрастанию ошибки прогноза. В данной работе используется Lasso-регрессия для выбора наилучшего набора экзогенных факторов.

 Для прогнозирования временных рядов использовались данные о концентрации углекислого газа () (CO2 concentration) в Тверской области, а в качестве экзогенных переменных рассматривались:

* + - * Температура на уровне 30 метров () (T\_30m);
			* Относительная влажность воздуха на уровне 50 метров (%) (Relative humidity at 50);
			* Солнечная радиация () (Solar radiation).

 Для оценки значимости признаков применялась Lasso-регрессия:

 (1)

 Для построения прогноза использовалась модель SARIMAX:

 (2)

 не сопряжены с

 - наименьшие из возможных

 - оператор Лага

Результаты исследования показали, что применение Lasso-регрессии позволило выделить наиболее значимые переменные: температура и влажность. Исключение солнечной радиации из модели снизило ошибку прогноза и уменьшило риск переобучения:

| **Параметры: экзогенные переменные** | **Ошибка, RMSE** | **Ошибка\*, %** |
| --- | --- | --- |
| Без экзогенных переменных (модель SARIMA) | 17.50 | 4.23 |
| T\_30m | 17.40 | 4.20 |
| Relative humidity at 50m | 15.96 | 3.85 |
| Solar radiation | 21.98 | 5.31 |
| T\_30m + Relative humidity at 50m | 17.02 | 4.11 |
| T\_30m + Solar radiation | 20.17 | 4.87 |
| Relative humidity at 50m + Solar radiation | 20.05 | 4.84 |
| T\_30m + Relative humidity at 50m + Solar radiation | 19.42 | 4.69 |

Таблица 1: Значения RMSE для различных наборов экзогенных переменных в модели SARIMAX

\* Ошибка - RMSE/mean, где mean(CO2 concentration) = 414. 01 за весь временной период.

Использование Lasso-регрессии для отбора экзогенных переменных в модели SARIMAX доказало свою эффективность. Полученные результаты могут быть полезны в различных областях науки, где требуется анализ временных рядов.

**Литература**

1. Hyndman R. J., Athanasopoulos G., 2018: Forecasting: Principles and Practice, Monash University.
2. Montgomery D. C., Jennings C. L., Kulahci M., 2015: Time Series Analysis and Forecasting, Wiley.