МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

Федеральное государственное бюджетное образовательное

учреждение высшего образования

«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»



Факультет прикладной математики и информатики

Кафедра ТПИ

Дисциплина: «Методы и технологии анализа больших данных»

Лабораторная работа №1

**ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ**

Факультет: ФПМИ

Группа: ПМИМ-31

Выполнили: Монгуш Н. С., Тарулин М. А., Филипенко Ю. Д.

Преподаватель: Сивак М. А.

Дата выполнения:

Отметка о защите:

**Задание**

1. Получить общую характеристику данных: значения основных статистик, значимость признаков, оценка параметров распределения признаков, оценка взаимодействия признаков друг с другом.
2. Построить различные графики для визуального анализа данных.
3. Получить информацию о наличии пропусков в данных, при их наличии предложить способы избавления от них.
4. Получить информация о наличии нетипичных наблюдений (выбросов) в данных. При наличии таких наблюдений предложить варианты работы с ними.

**Исходные данные**

Исходные данные представляют собой совокупность эколого-энергетических параметров сжигания жидких углеводородов в присутствии различных распылителей-разбавителей, а именно состав промежуточных и конечных продуктов сгорания для 8 сочетаний топлив и разбавителей: сырая нефть-пар, дизельное топливо-воздух, дизельное топливо-пар, мазут-воздух, мазут-пар, керосин-воздух, керосин-пар, отработанное масло-пар, полученные в результате аппроксимации экспериментальных данных и на каждое сочетание приходится 10 тысяч наблюдений. Визуальным представлением этих данных являются топливные карты, состоящие из линий уровня компонента дымовых газов, по оси абсцисс находится расход топлива, по оси ординат – расход вводимого в пламя компонента (Рис. 1.).

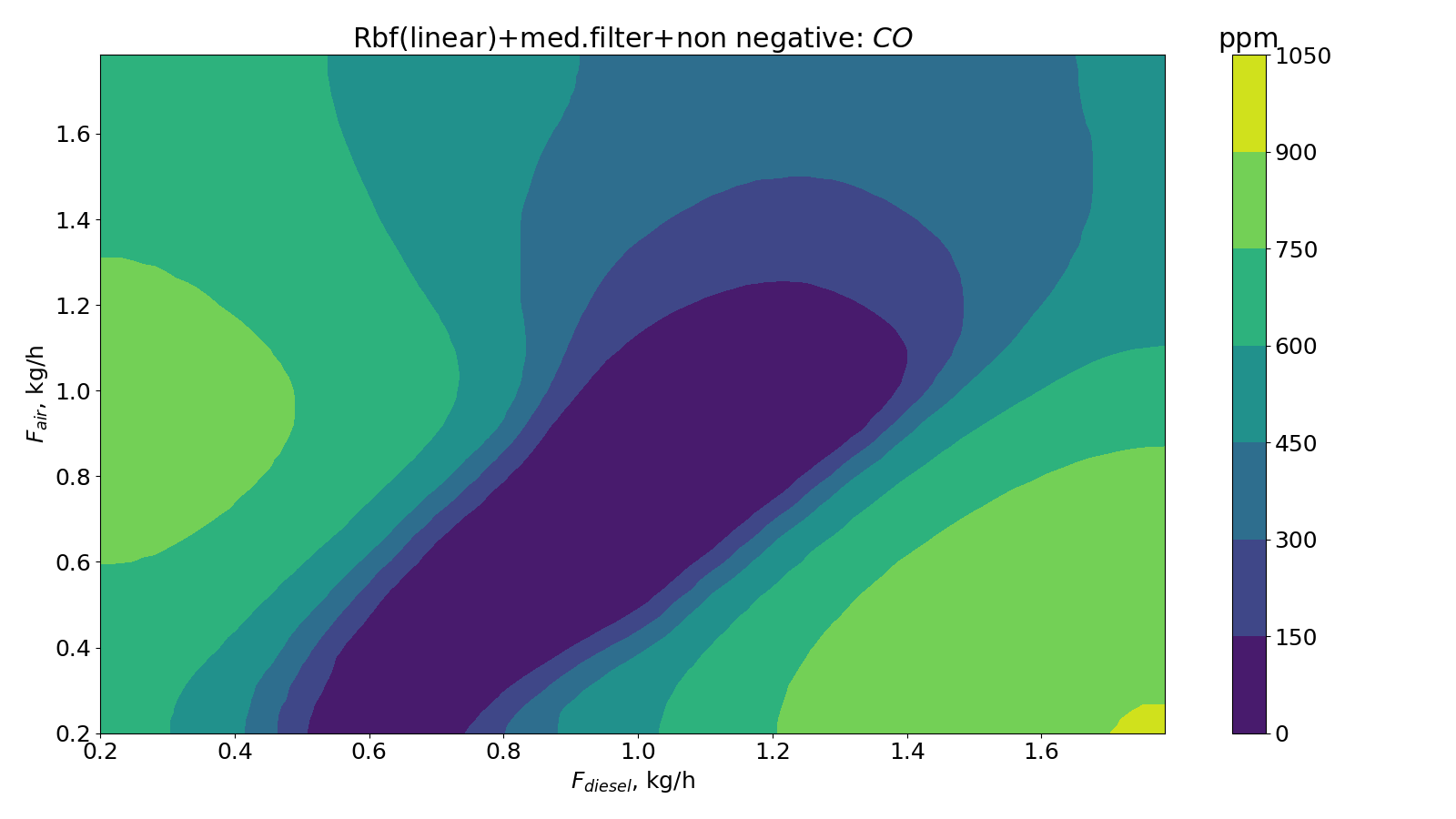


Рис. 1. Пример карты режима горения для дизельного топлива и вводимого в пламя воздуха для CO.

Таблица 1. Пример исходных данных для сырой нефти и вводимого в пламя пара.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F\_fuel | F\_steam | O2 | CO | NO | NO2 | NO­­x | CO­­2 | SO­2 |
| **0** | 0.200 | 1.784 | 10.924389 | 1436.467675 | 85.878288 | 3.08404 | 88.968459 | 7.820647 | 155.282350 |
| **1** | 0.216 | 1.784 | 10.924389 | 1436.467675 | 85.878288 | 3.08404 | 88.968459 | 7.820647 | 155.282350 |
| **2** | 0.232 | 1.784 | 10.924389 | 1436.311000 | 85.878288 | 3.08404 | 88.968459 | 7.820647 | 155.282350 |
| **3** | 0.248 | 1.784 | 10.918877 | 1435.131099 | 85.878288 | 3.08404 | 88.968459 | 7.821372 | 155.282350 |
| **4** | 0.264 | 1.784 | 10.915218 | 1433.974051 | 85.887051 | 3.08404 | 88.971090 | 7.825922 | 155.340263 |

**Оценка пропущенных наблюдений и five-summary statistics**

В связи с особенностями построения аппроксимированных поверхностей в исходных данных отсутствуют пропущенные наблюдения, перейдем к анализу five-summary statistics:

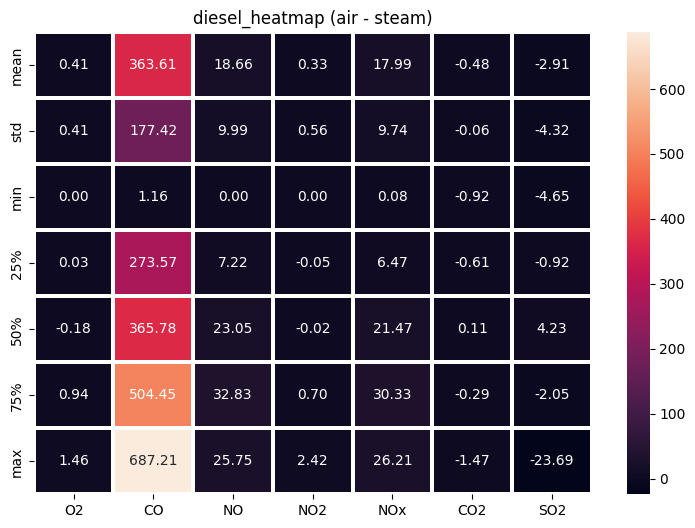


Рис.2. Результат вычитания двух матриц five-summary statistics c вводимым воздухом и паром для дизельного топлива.

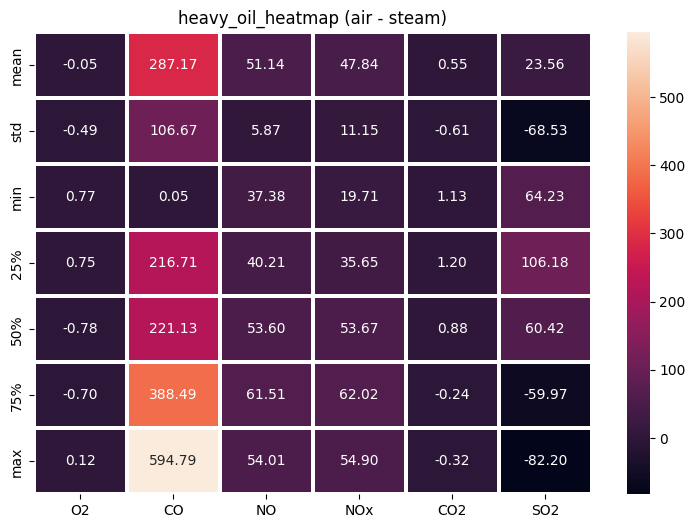


Рис.3. Результат вычитания двух матриц five-summary statistics c вводимым воздухом и паром для мазута.

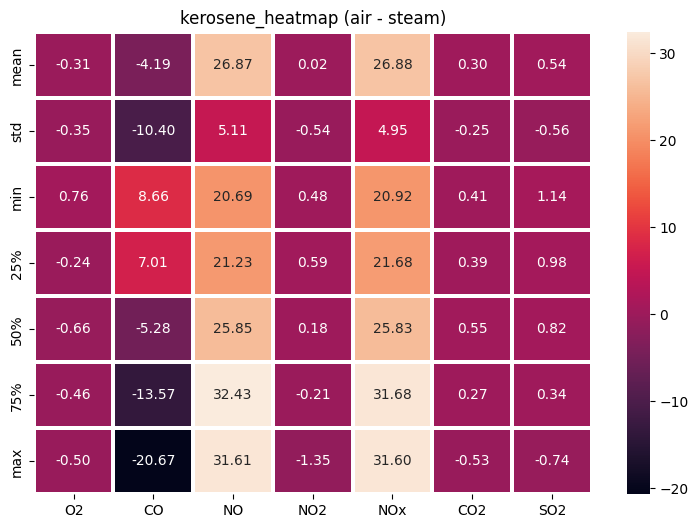


Рис.4. Результат вычитания двух матриц five-summary statistics c вводимым воздухом и паром для керосина.

Из рис. 2-4 можно увидеть, что у матриц five-summary statistics с вводимым в пламя паром в значительной степени ниже концентрации CO, NOx и NO в дымовых газах во всех исследуемых топливах, за исключением керосина, где наблюдается обратная зависимость.

**Визуальный анализ данных**

Начнем визуальный анализ данных с графиков boxplot – графиков, которые отражают форму распределения, медиану, квартили и выбросы.

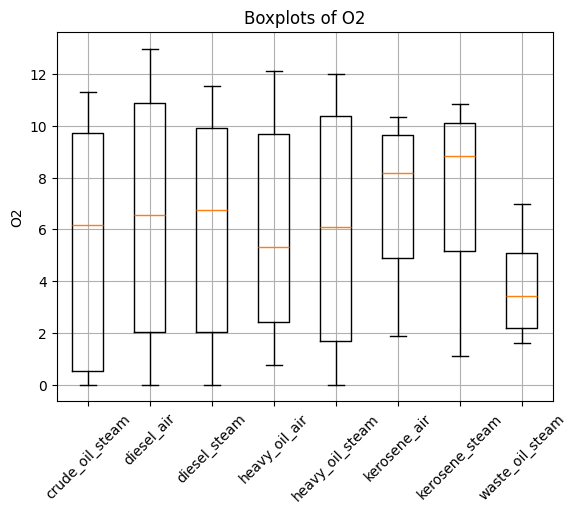


Рис. 2. График boxplot для компонента O2.

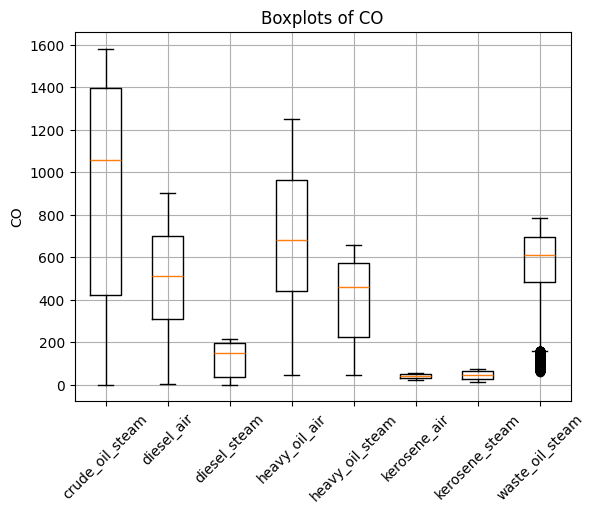


Рис. 3. График boxplot для компонента CO.

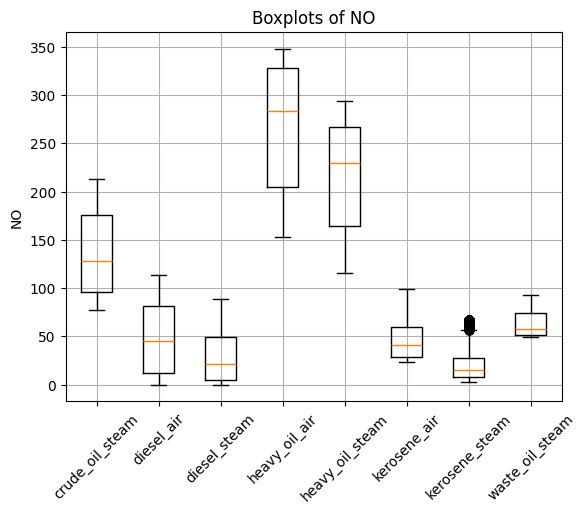


Рис. 4. График boxplot для компонента NO.

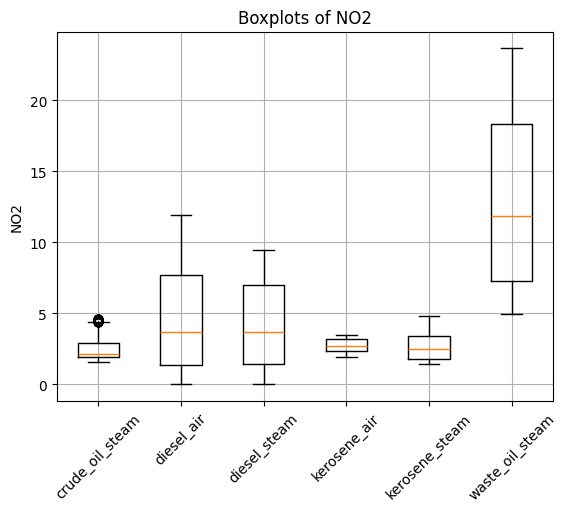


Рис. 5. График boxplot для компонента NO­2.

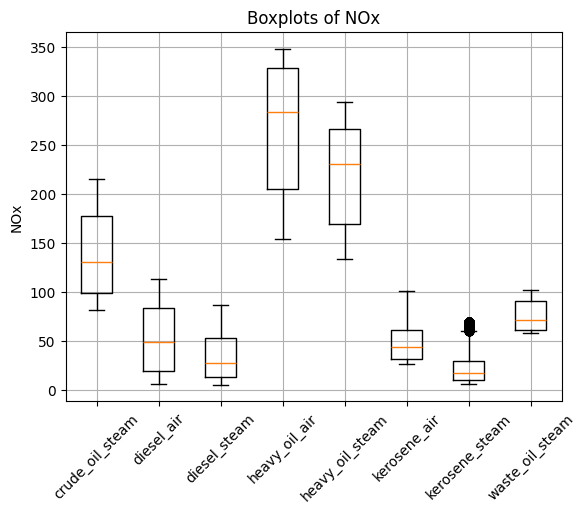


Рис. 6. График boxplot для компонента NO­x.

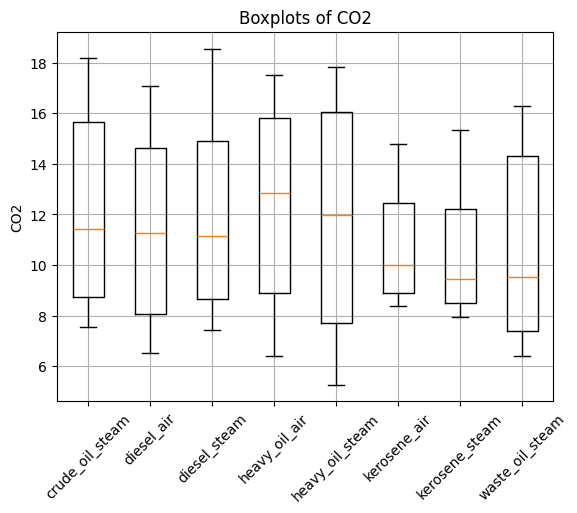


Рис. 7. График boxplot для компонента CO2.

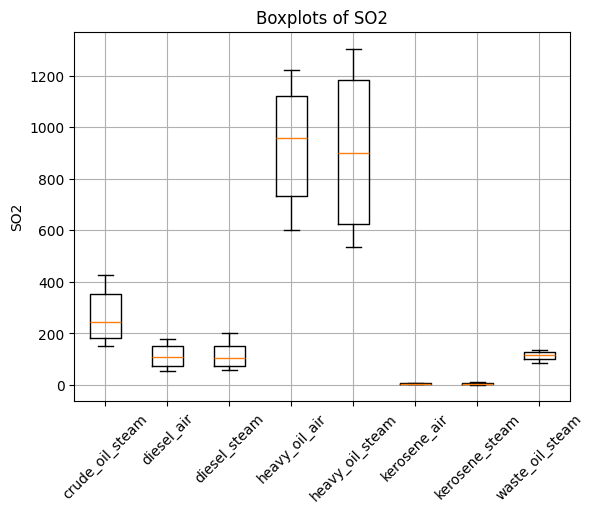


Рис. 8. График boxplot для компонента SO­2.

По представленным графикам видно, что выбросы присутствуют на графике 3, 4, 5, 6. Как можно увидеть из рис. 9-12 выбросы, отмеченные на графиках boxplot, присутствуют и в исходных данных, которые формировали опорные точки для интерполяционной поверхности. Данные выбросы можно объяснить тем, что экспериментальные данные имеют малую дискретность относительно расхода топлива и вводимого компонента.

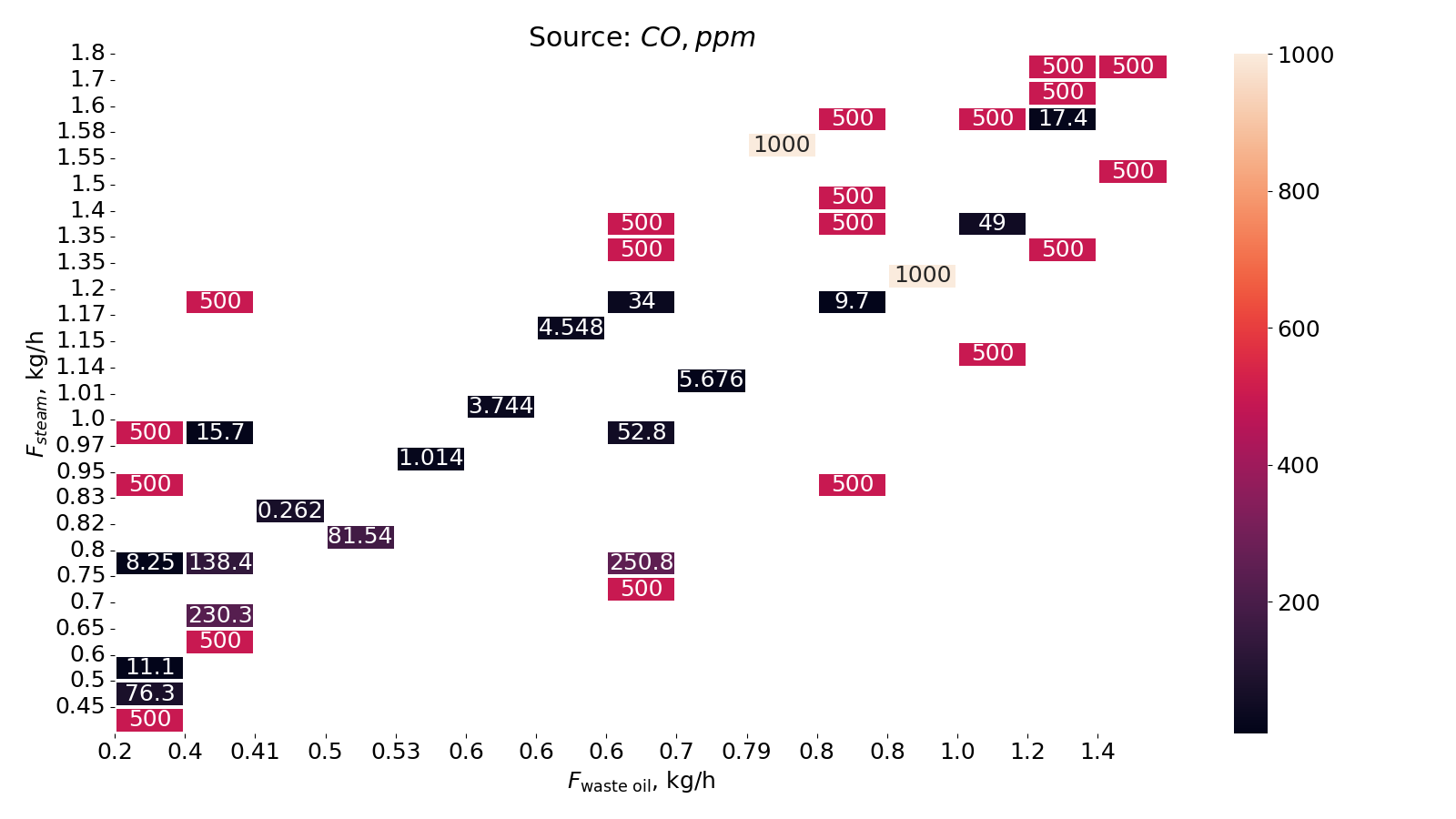


Рис. 9. Опорные точки компонента CO для отработанного масла и вводимого пара.

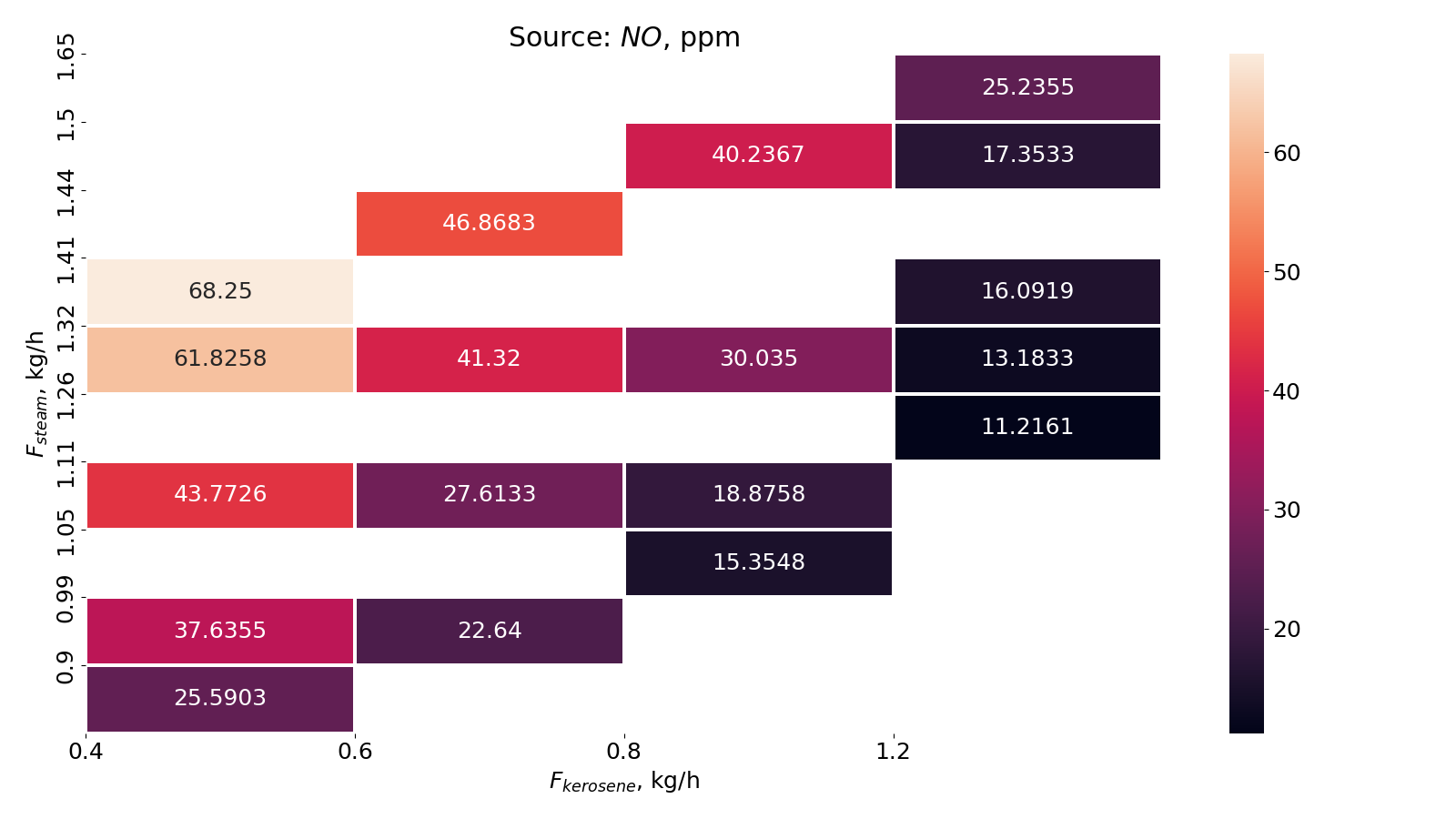


Рис. 10. Опорные точки компонента NO для керосина и вводимого пара.

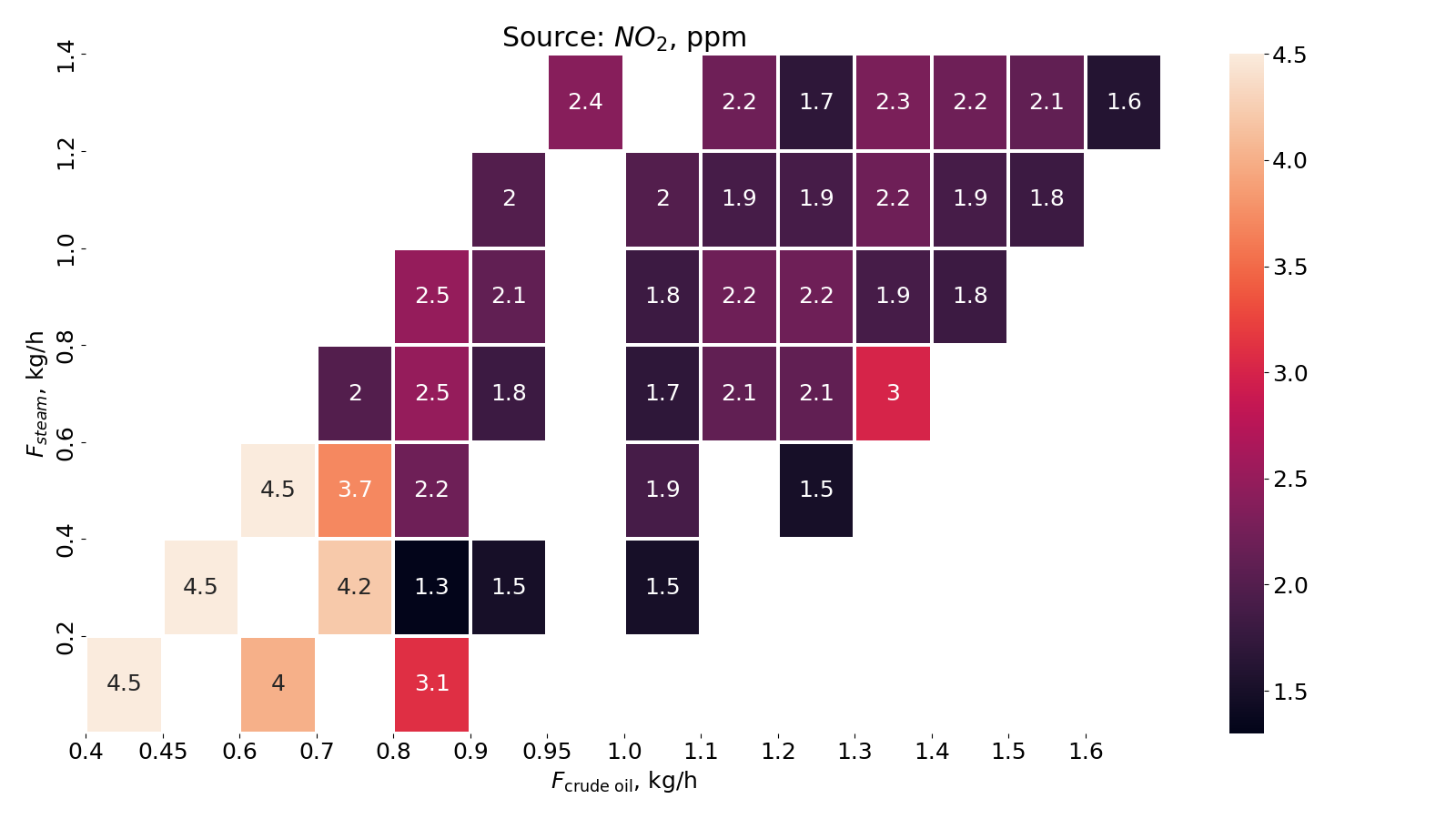


Рис. 11. Опорные точки компонента NO2 для керосина и вводимого пара.

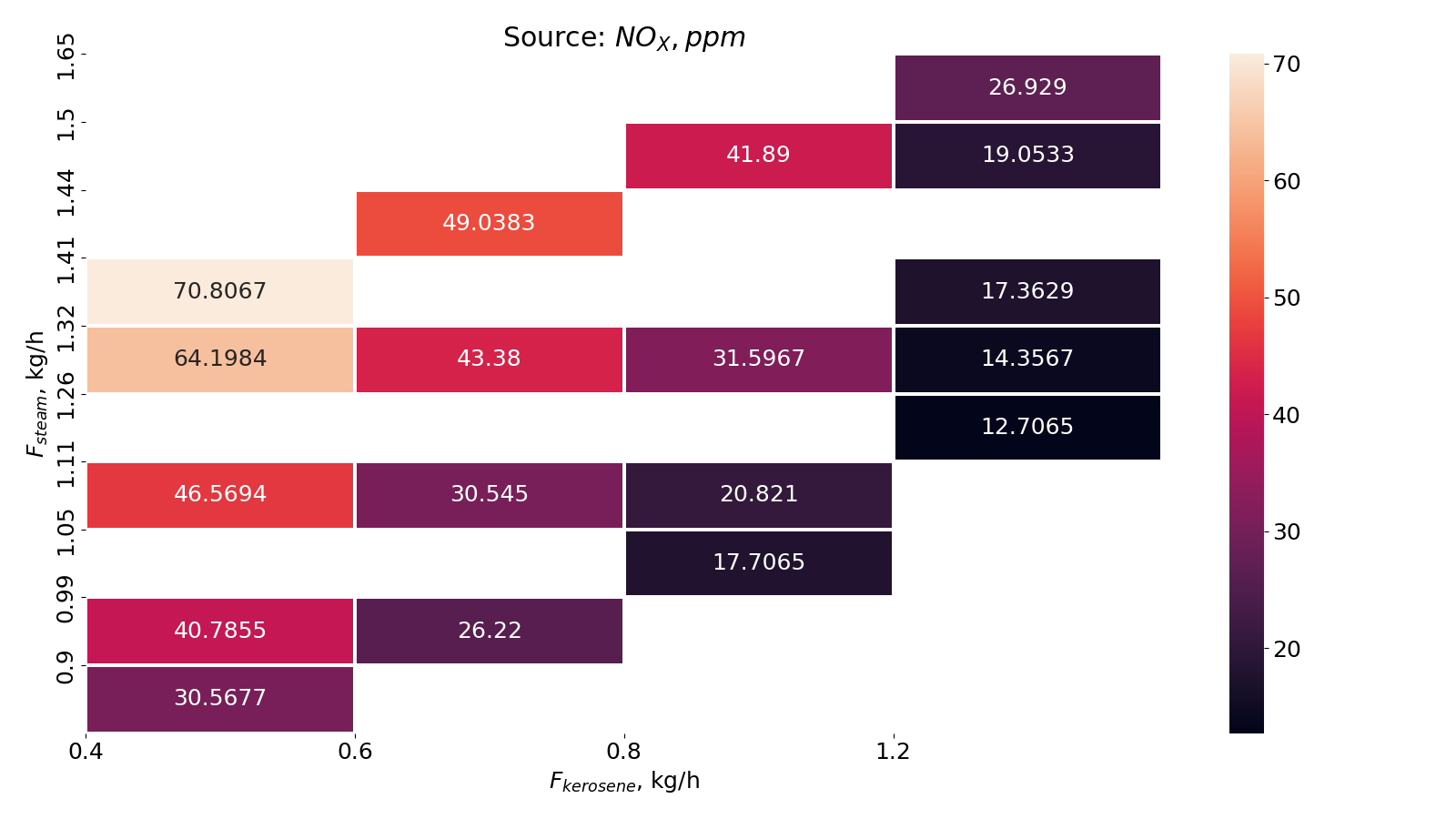


Рис. 12. Опорные точки компонента NOx для керосина и вводимого пара.

Далее рассмотрим графики парных сеток. Графики парных сеток – это инструмент визуализации данных, который отображает все возможные пары переменных в наборе данных в виде матрицы графиков, которые показывают взаимосвязь между парами переменных, с помощью этих графиков можно оценить корреляцию и зависимости между переменными.

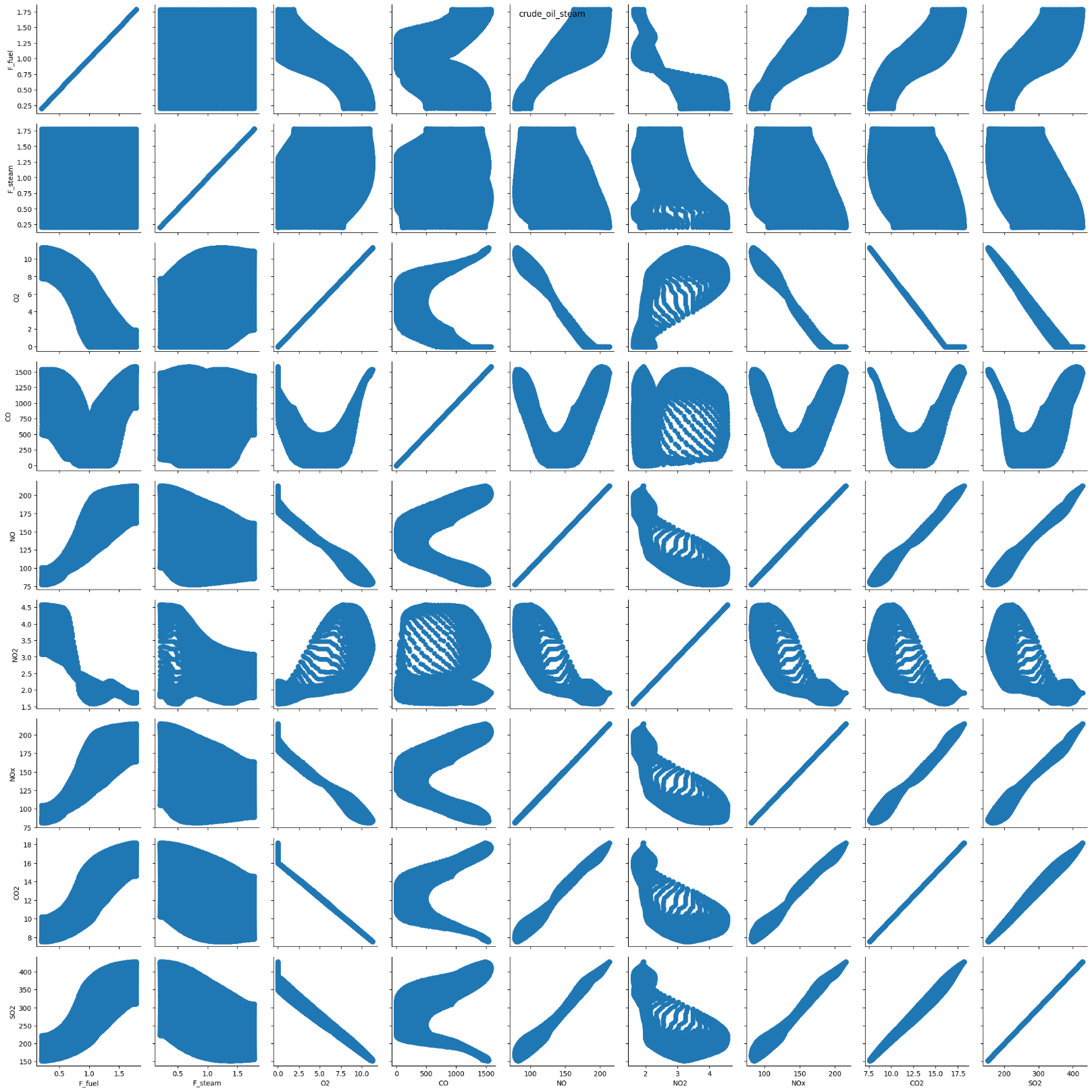


Рис. 13. График парных сеток для сырой нефти и вводимого в пламя пара.

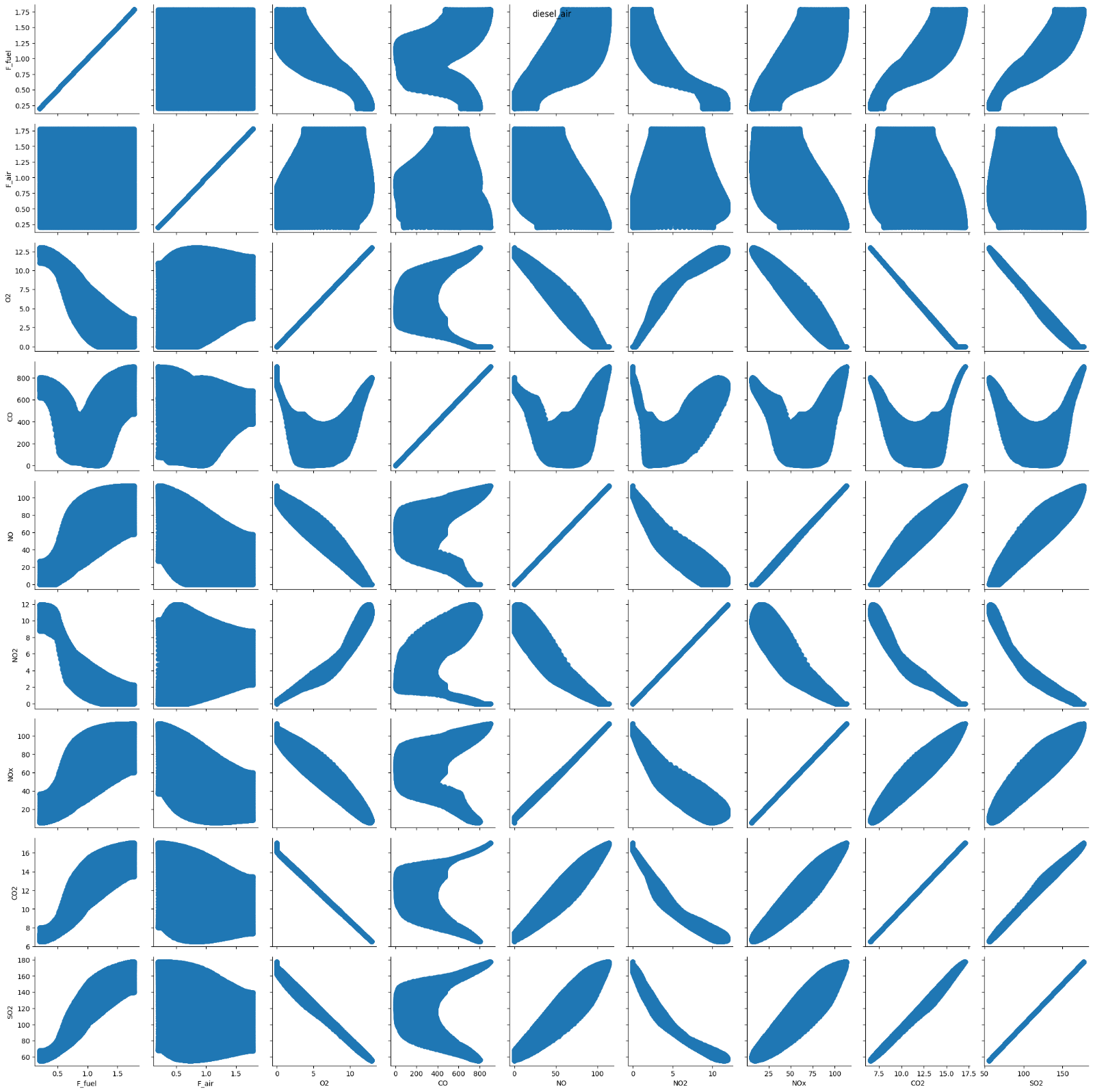


Рис. 14. График парных сеток для дизельного топлива и вводимого в пламя воздуха.

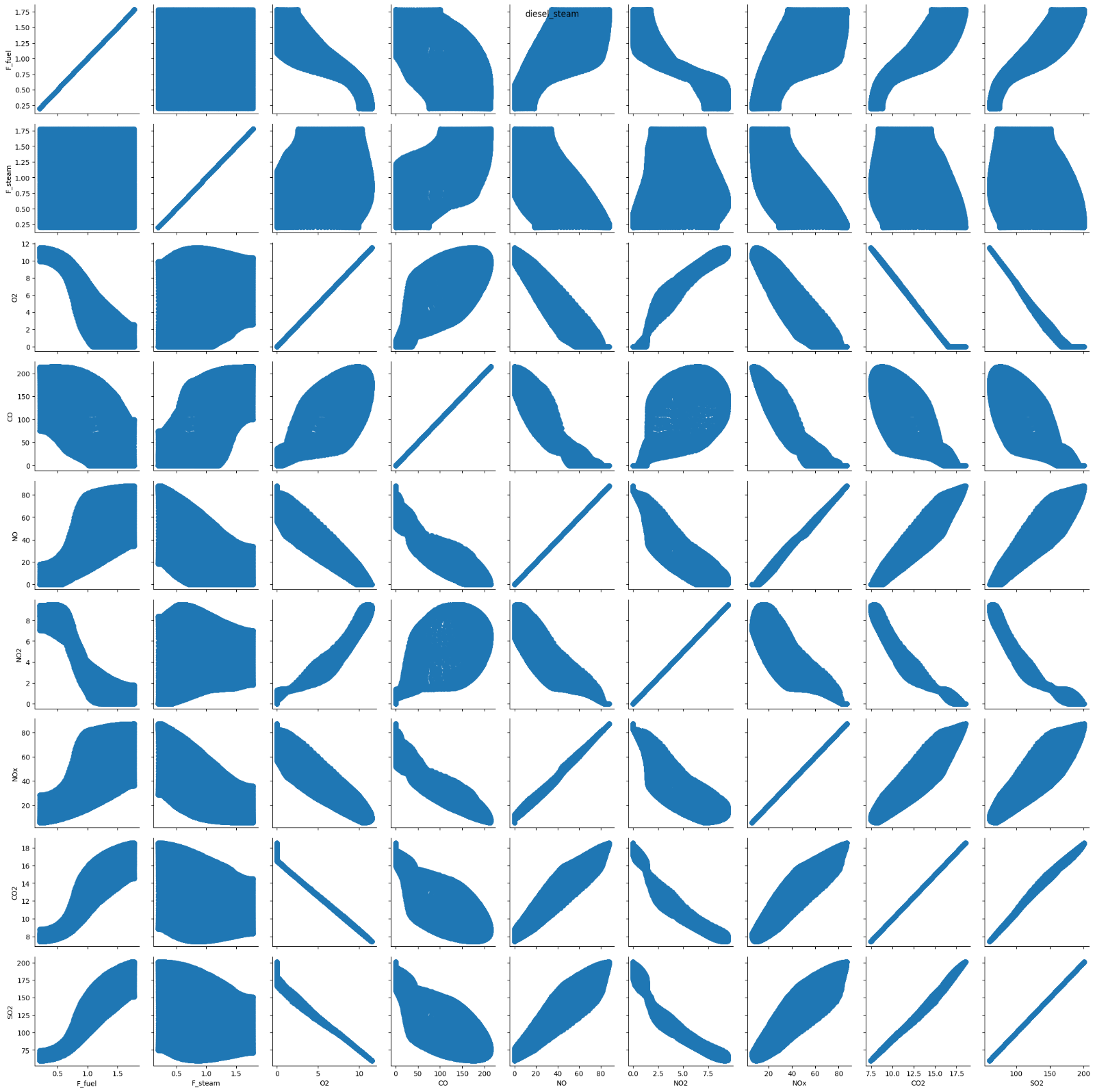


Рис. 15. График парных сеток для дизельного топлива и вводимого в пламя пара.

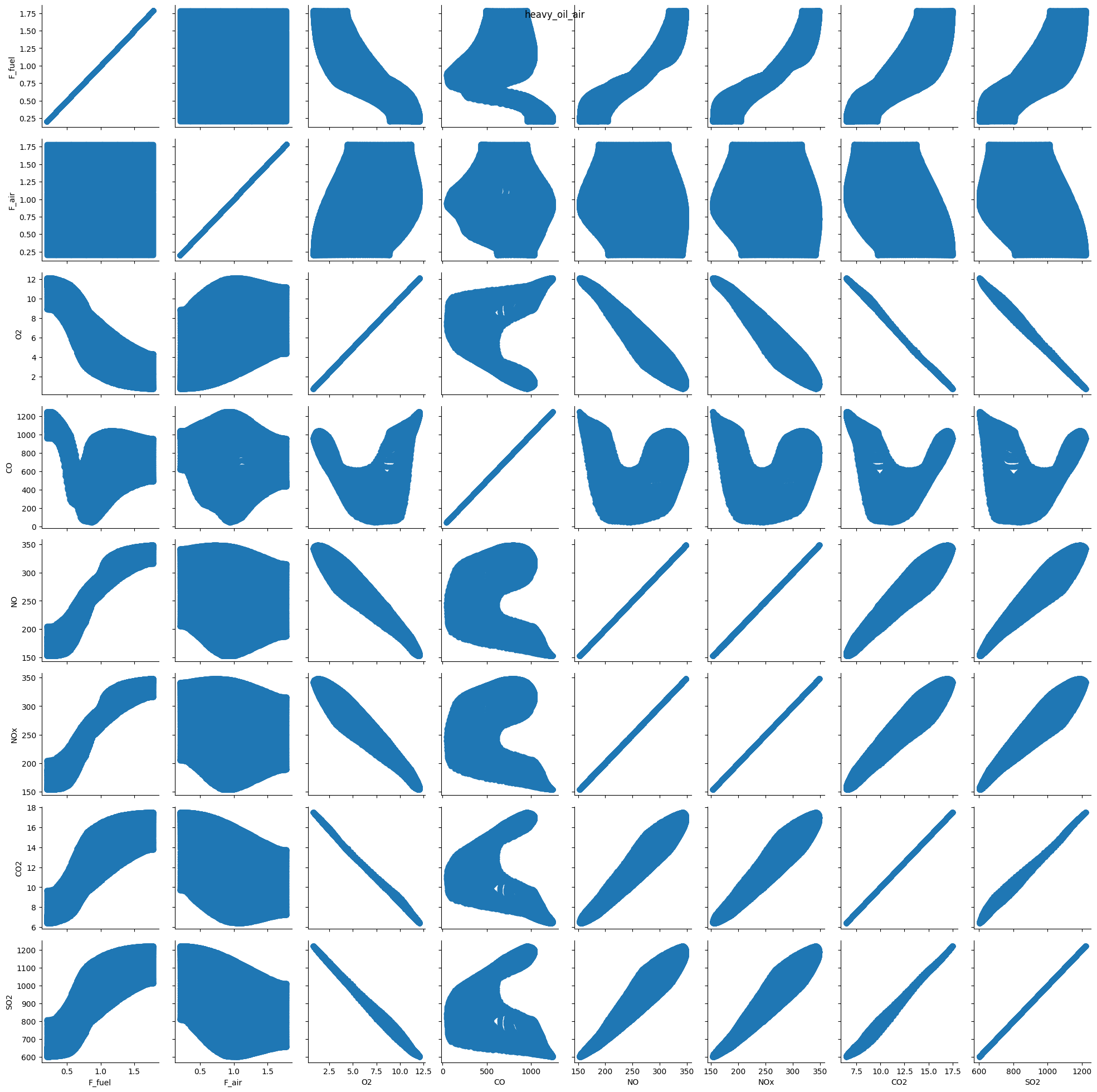


Рис. 16. График парных сеток для мазута и вводимого в пламя воздуха.

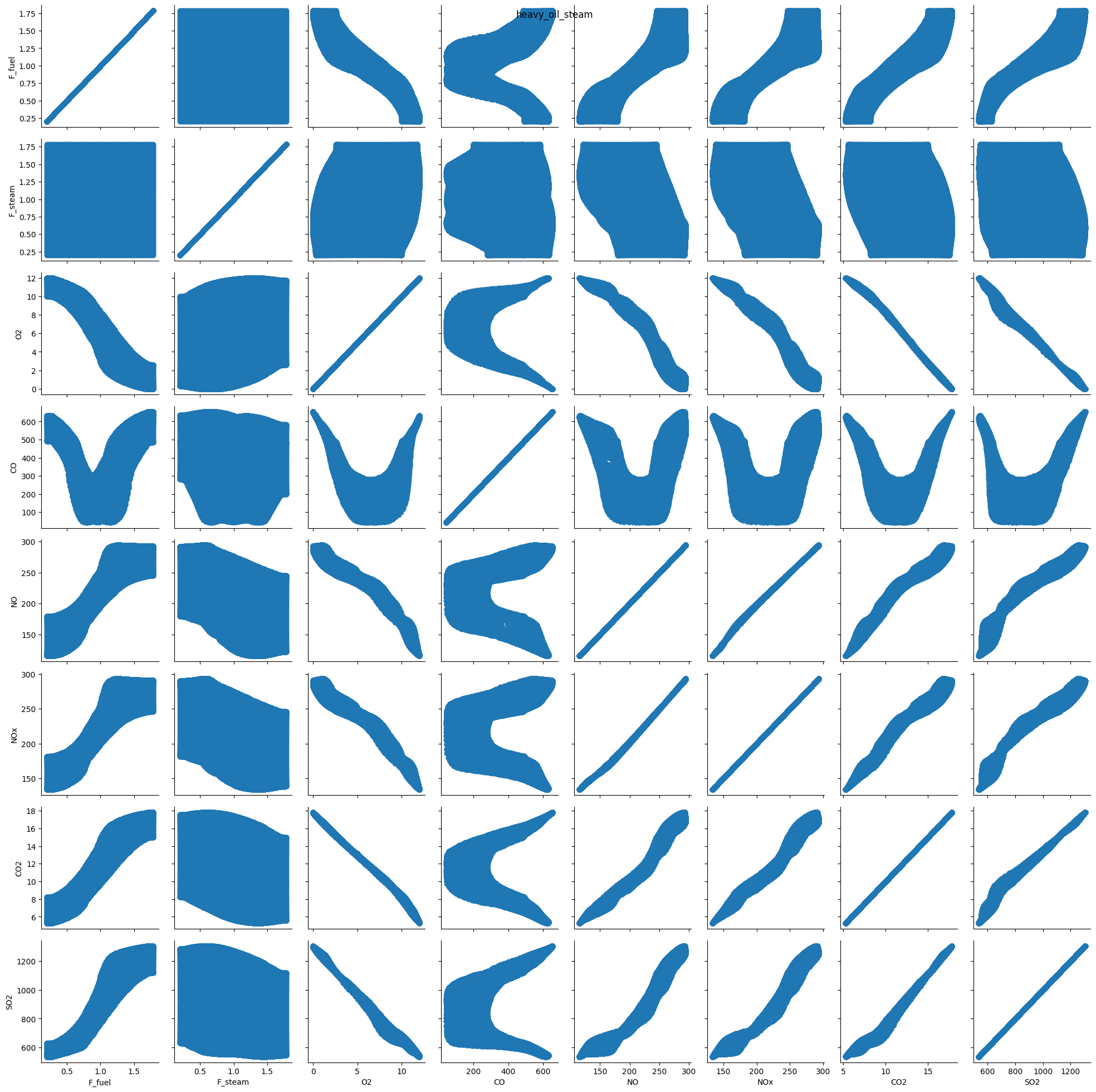


Рис. 17. График парных сеток для мазута и вводимого в пламя пара.

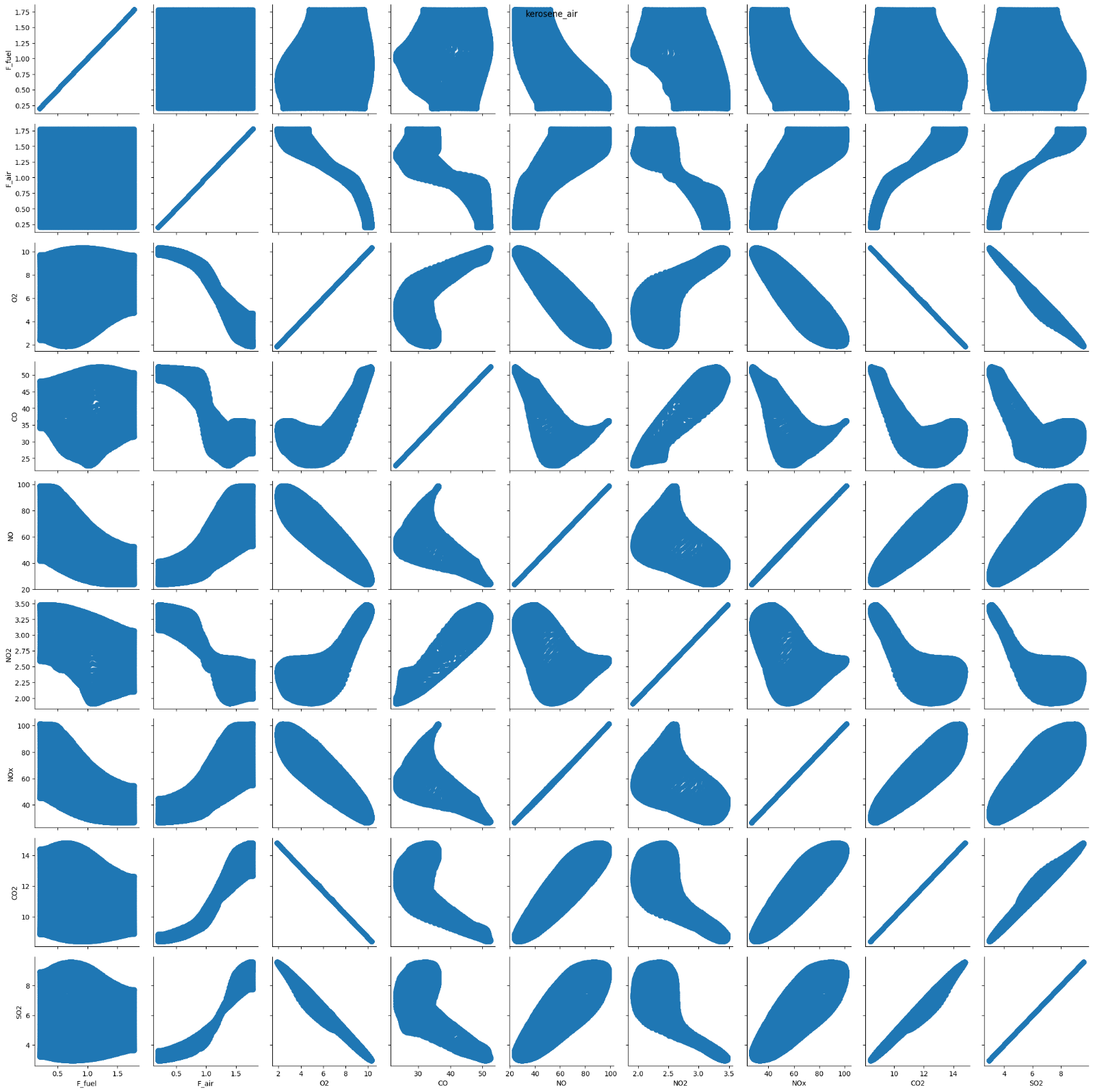


Рис. 18. График парных сеток для керосина и вводимого в пламя воздуха.

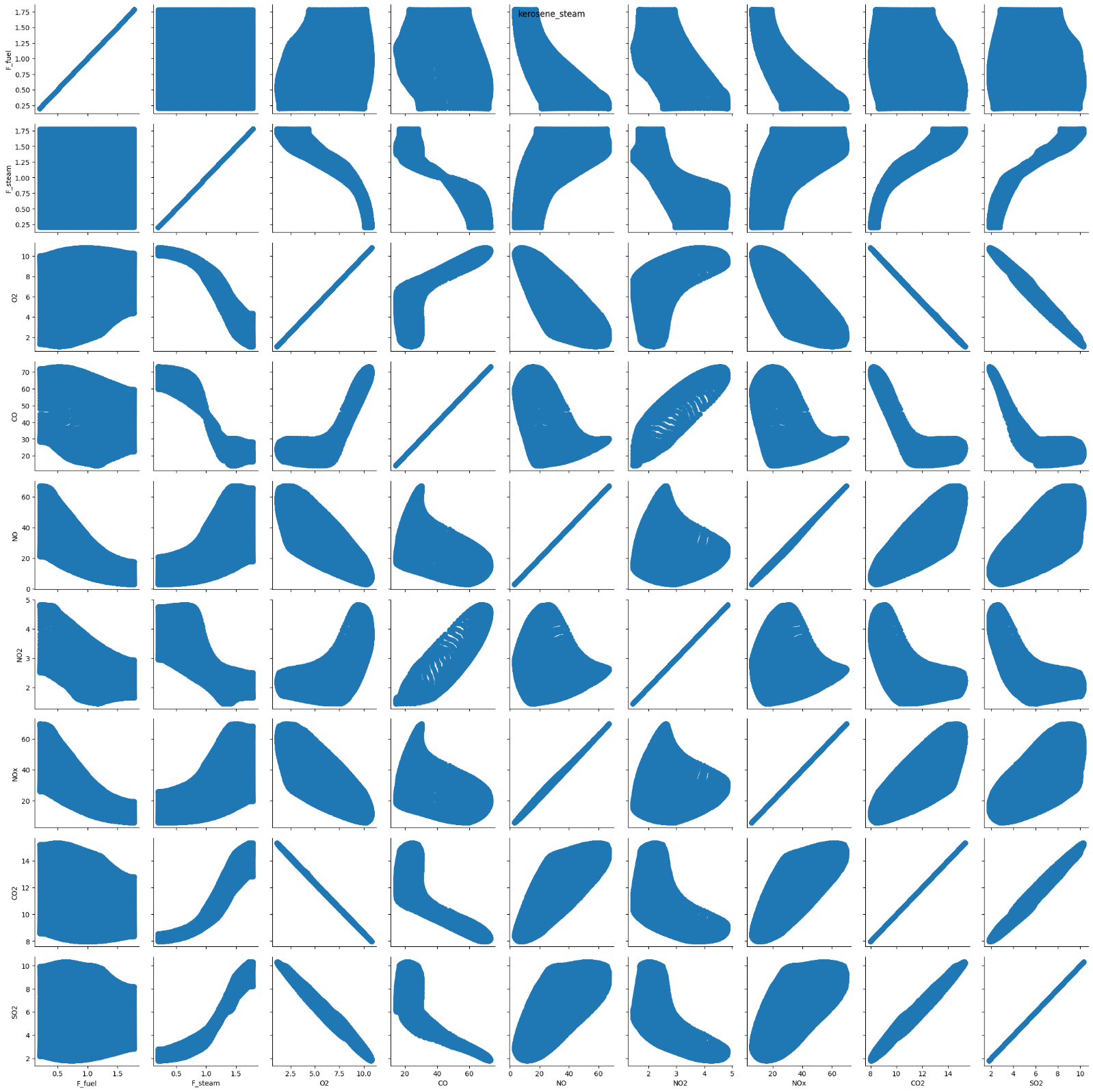


Рис. 19. График парных сеток для керосина и вводимого в пламя пара.

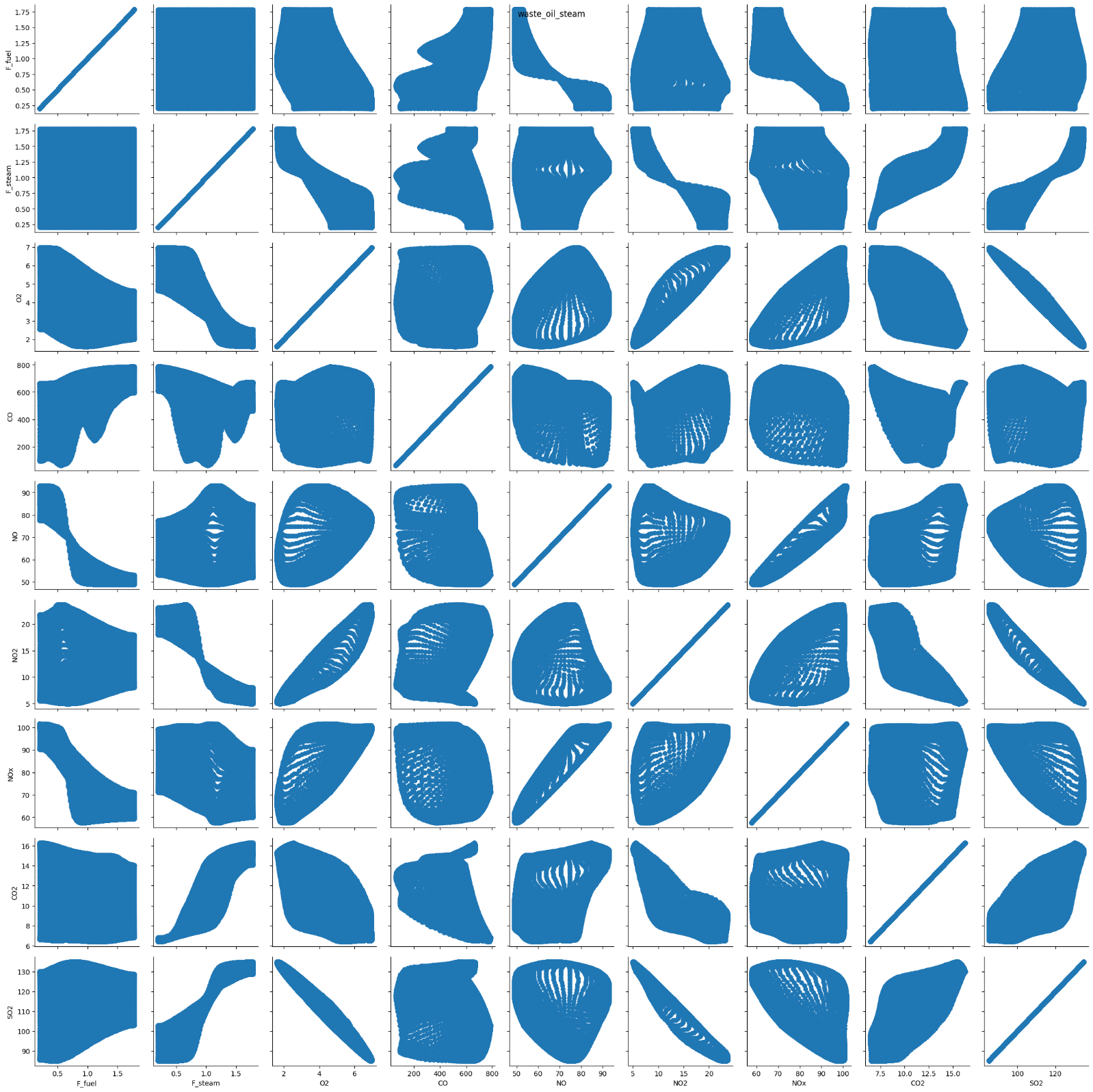


Рис. 20. График парных сеток для отработанного масла и вводимого в пламя пара.

Помимо очевидных линейных зависимостей, которые будут отражены далее в корреляционной матрице более наглядным образом, можно отметить параболический характер распространения для компонента CO для сочетаний: сырая нефть-пар (рис. 13), дизель-воздух (рис. 14), мазут-воздух (рис. 16), мазут-пар (рис. 17).

Корреляционная матрица представляет собой таблицу, в которой коэффициенты корреляции между парами переменных измеряются и отображаются. Коэффициент корреляции измеряет степень линейной взаимосвязи между двумя переменными. Он может принимать значения от -1 до 1:

Значение 1 означает положительную линейную корреляцию, что означает, что переменные изменяются вместе в одном направлении.

Значение -1 означает отрицательную линейную корреляцию, что означает, что переменные изменяются в противоположных направлениях.

Значение 0 означает отсутствие линейной корреляции между переменными.

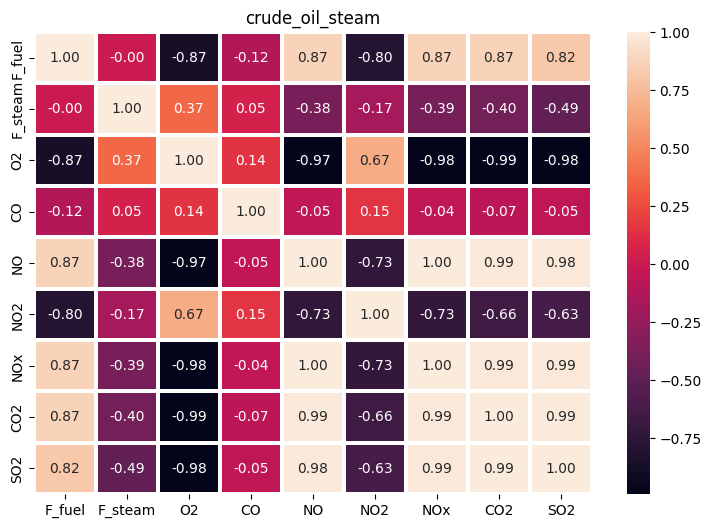


Рис. 21. Корреляционная матрица для сырой нефти и вводимого в пламя пара.

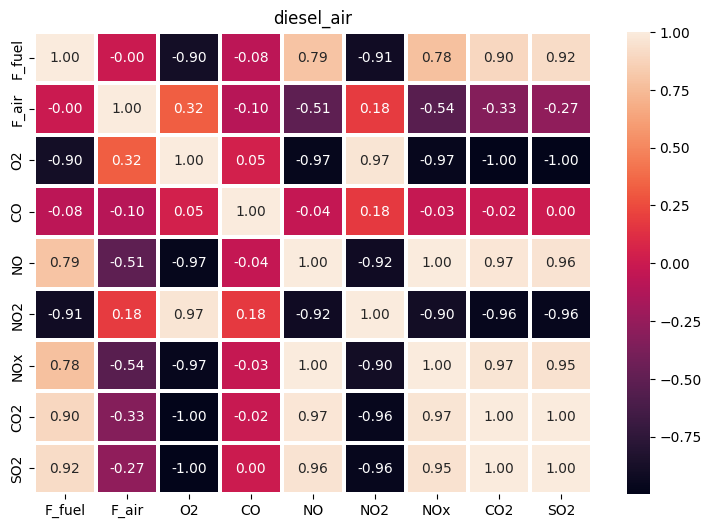


Рис. 22. Корреляционная матрица для дизельного топлива и вводимого в пламя воздуха.

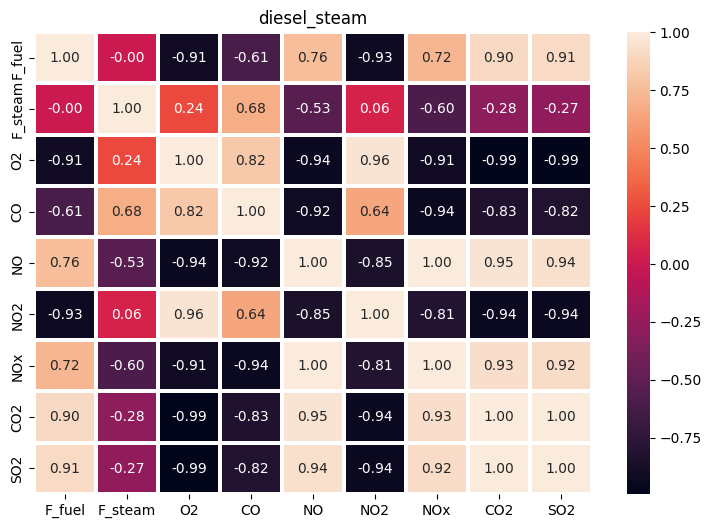


Рис. 23. Корреляционная матрица для дизельного топлива и вводимого в пламя пара.

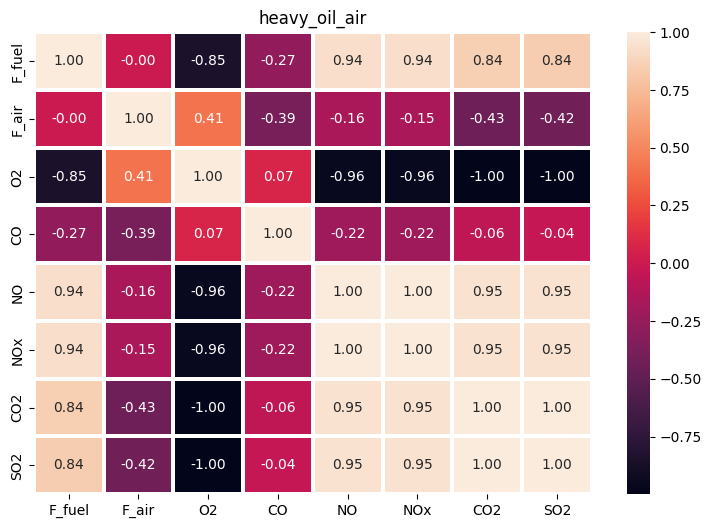


Рис. 24. Корреляционная матрица для мазута и вводимого в пламя воздуха.

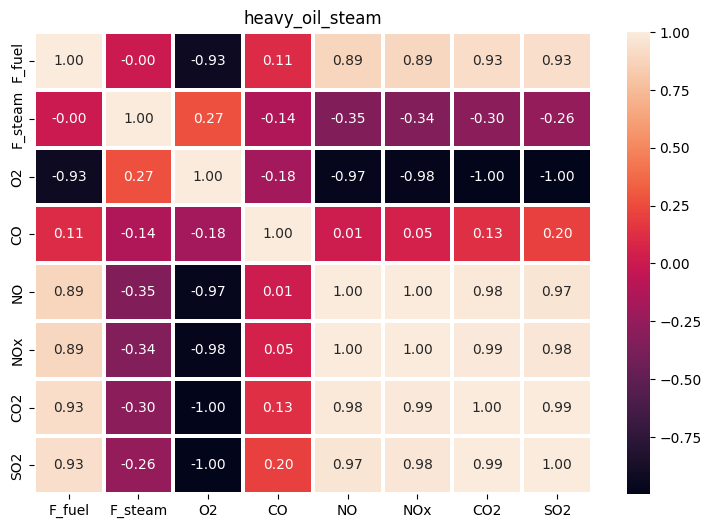


Рис. 25. Корреляционная матрица для мазута и вводимого в пламя пара.

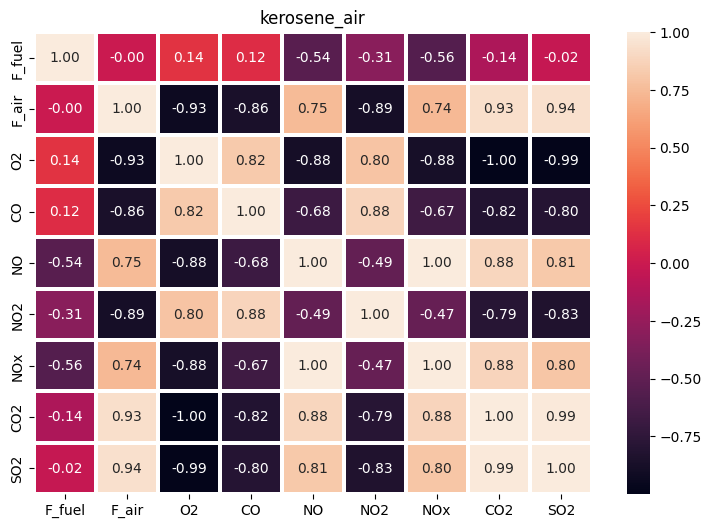
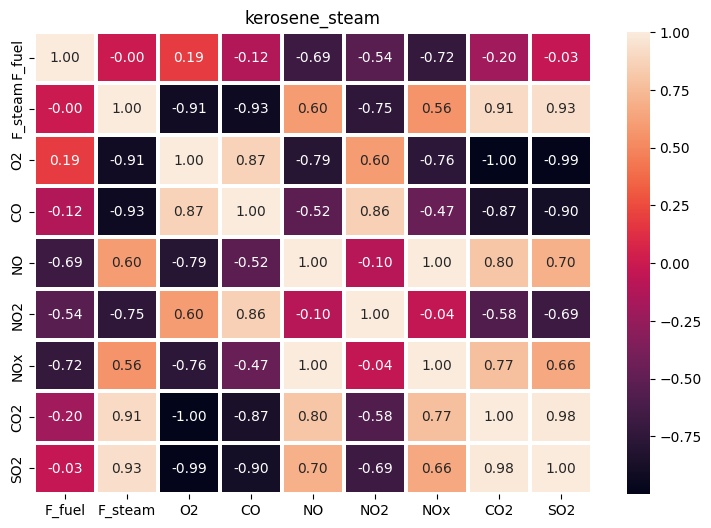
Рис. 26. Корреляционная матрица для мазута и вводимого в пламя воздуха.

Рис. 27. Корреляционная матрица для керосина и вводимого в пламя пара.

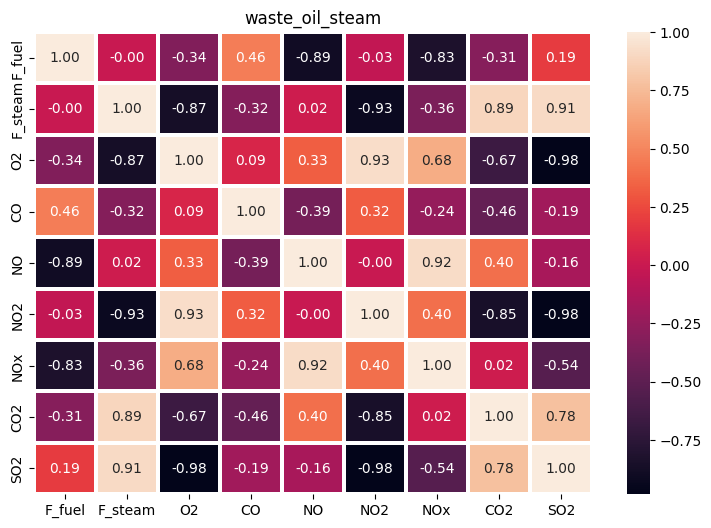


Рис. 28. Корреляционная матрица для отработанного масла и вводимого в пламя пара.