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.).

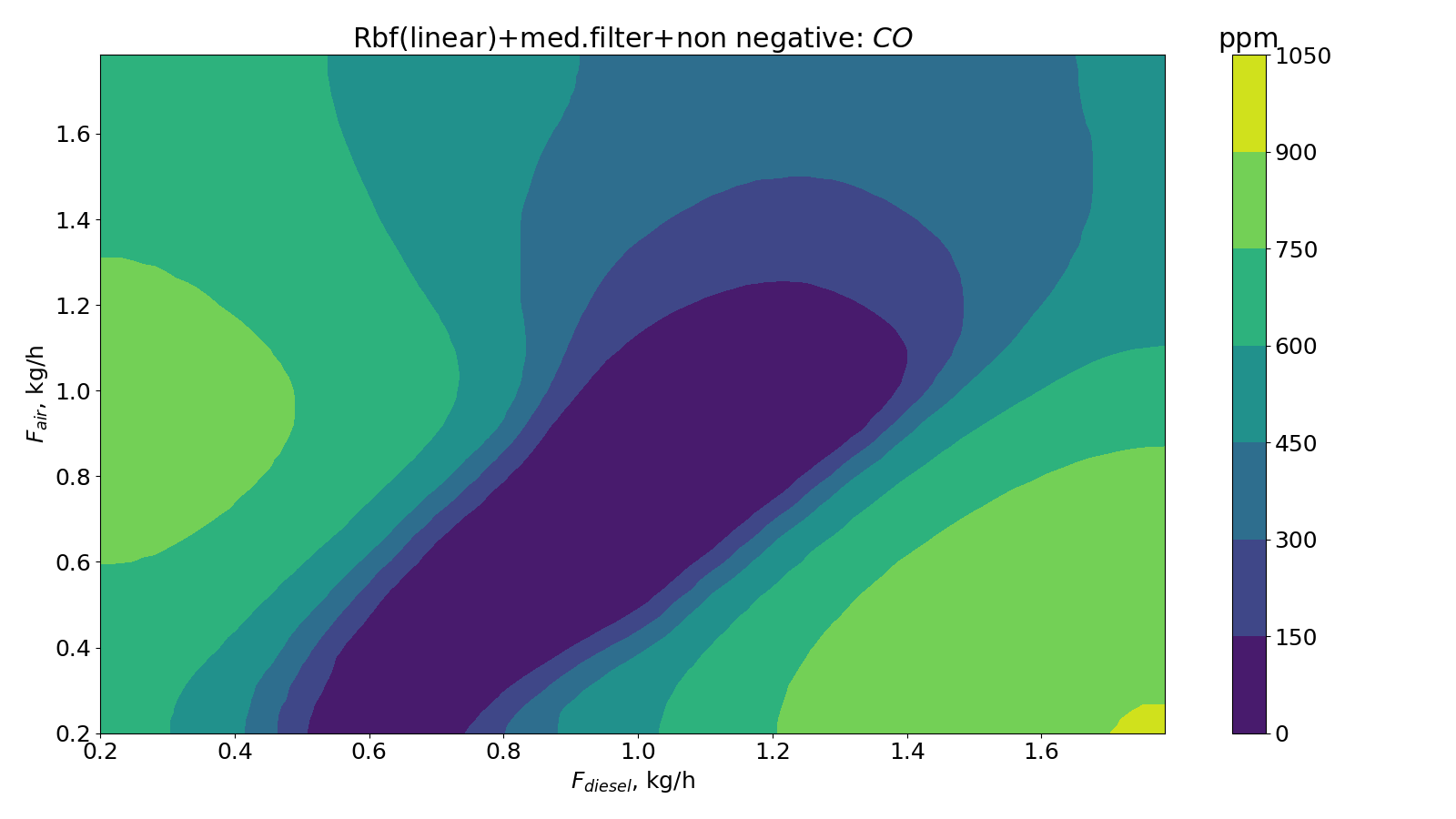


Рис. 1. Пример карты режима горения для дизельного топлива и вводимого в пламя воздуха для CO.

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:

Таблица 2. five-summary statistics для сочетания сырая нефть-пар.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F\_fuel | F\_steam | O2 | CO | NO | NO2 | NO­­x | CO­­2 | SO­2 |
| **count** | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| **mean** | 0.9920 | 0.9920 | 5.497351 | 902.016513 | 136.19105 | 2.50441 | 138.69798 | 12.14760 | 267.76652 |
| **std** | 0.4618 | 0.4618 | 4.281942 | 537.389760 | 43.228927 | 0.80729 | 42.639550 | 3.544979 | 91.156857 |
| **min** | 0.2000 | 0.2000 | 0.000000 | 0.000000 | 77.521582 | 1.59144 | 81.536263 | 7.554174 | 152.14599 |
| **25%** | 0.5960 | 0.5960 | 0.522354 | 421.443792 | 95.825941 | 1.91276 | 99.086982 | 8.721034 | 182.35649 |
| **50%** | 0.9920 | 0.9920 | 6.167016 | 1058.97781 | 128.34502 | 2.15704 | 130.37852 | 11.43766 | 245.07411 |
| **75%** | 1.3880 | 1.3880 | 9.741279 | 1397.31404 | 175.22244 | 2.91848 | 177.22921 | 15.64013 | 352.73310 |
| **max** | 1.784 | 1.78400 | 11.309021 | 1581.430033 | 213.135621 | 4.577867 | 215.065716 | 18.188900 | 427.542054 |

Таблица 3. five-summary statistics для сочетания дизельное топливо-воздух.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F\_fuel | F\_steam | O2 | CO | NO | NO2 | NO­­x | CO­­2 | SO­2 |
| **count** | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| **mean** | 0.99200 | 0.99200 | 6.421306 | 484.542919 | 48.086647 | 4.545273 | 52.411160 | 11.427592 | 112.554569 |
| **std** | 0.46188 | 0.46188 | 4.490491 | 257.662436 | 37.493270 | 3.581094 | 34.453863 | 3.411426 | 39.544774 |
| **min** | 0.20000 | 0.20000 | 0.000000 | 1.162475 | 0.000000 | 0.000000 | 5.669769 | 6.509279 | 55.385421 |
| **25%** | 0.59600 | 0.59600 | 2.045992 | 307.969935 | 11.797590 | 1.381264 | 19.710568 | 8.067942 | 74.174307 |
| **50%** | 0.99200 | 0.99200 | 6.559653 | 512.791778 | 44.744743 | 3.695360 | 48.553762 | 11.274453 | 109.430333 |
| **75%** | 1.38800 | 1.38800 | 10.871679 | 701.007658 | 81.804441 | 7.701282 | 83.278519 | 14.621373 | 149.779775 |
| **max** | 1.78400 | 1.78400 | 12.987349 | 901.410260 | 113.894443 | 11.908415 | 113.216933 | 17.087654 | 177.272712 |

Таблица 4. five-summary statistics для сочетания дизельное топливо-пар.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F\_fuel | F\_steam | O2 | CO | NO | NO2 | NO­­x | CO­­2 | SO­2 |
| **count** | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| **mean** | 0.99200 | 0.99200 | 6.015459 | 120.934151 | 29.427570 | 4.215731 | 34.424735 | 11.908048 | 115.469180 |
| **std** | 0.46188 | 0.46188 | 4.084898 | 80.241320 | 27.507953 | 3.018783 | 24.713541 | 3.472821 | 43.863793 |
| **min** | 0.20000 | 0.20000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 5.594444 | 7.431294 | 60.031310 |
| **25%** | 0.59600 | 0.59600 | 2.019518 | 34.396323 | 4.582426 | 1.427811 | 13.241146 | 8.677754 | 75.092298 |
| **50%** | 0.99200 | 0.99200 | 6.736857 | 147.012339 | 21.690060 | 3.717729 | 27.080608 | 11.164907 | 105.197598 |
| **75%** | 1.38800 | 1.38800 | 9.930256 | 196.556571 | 48.973120 | 7.003432 | 52.953452 | 14.914366 | 151.825230 |
| **max** | 1.78400 | 1.78400 | 11.528793 | 214.202191 | 88.144600 | 9.493099 | 87.007800 | 18.556423 | 200.966020 |

Таблица 5. five-summary statistics для сочетания мазут-воздух.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F\_fuel | F\_steam | O2 | CO | NO | NO­­x | CO­­2 | SO­2 |
| **count** | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| **mean** | 0.99200 | 0.99200 | 5.976554 | 685.621601 | 267.369860 | 267.708438 | 12.370760 | 930.646428 |
| **std** | 0.46188 | 0.46188 | 3.751934 | 302.147897 | 64.281553 | 63.973748 | 3.624875 | 203.508728 |
| **min** | 0.20000 | 0.20000 | 0.767464 | 43.826657 | 152.845653 | 153.617516 | 6.408058 | 601.206440 |
| **25%** | 0.59600 | 0.59600 | 2.432134 | 442.200219 | 204.952480 | 205.105966 | 8.908892 | 731.938346 |
| **50%** | 0.99200 | 0.99200 | 5.301761 | 679.765580 | 283.577378 | 283.906711 | 12.860010 | 959.597070 |
| **75%** | 1.38800 | 1.38800 | 9.693670 | 961.859601 | 328.083933 | 328.075175 | 15.820810 | 1122.337666 |
| **max** | 1.78400 | 1.78400 | 12.116081 | 1249.521890 | 348.252142 | 348.228552 | 17.501087 | 1222.563840 |

Таблица 6. five-summary statistics для сочетания мазут-пар.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F\_fuel | F\_steam | O2 | CO | NO | NO­­x | CO­­2 | SO­2 |
| **count** | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| **mean** | 0.99200 | 0.99200 | 6.030396 | 398.455134 | 216.230464 | 219.866489 | 11.822812 | 907.087650 |
| **std** | 0.46188 | 0.46188 | 4.241058 | 195.480268 | 58.415649 | 52.827420 | 4.235540 | 272.039773 |
| **min** | 0.20000 | 0.20000 | 0.000000 | 43.778546 | 115.467443 | 133.904130 | 5.280184 | 536.974350 |
| **25%** | 0.59600 | 0.59600 | 1.684322 | 225.486817 | 164.743703 | 169.460467 | 7.709826 | 625.762153 |
| **50%** | 0.99200 | 0.99200 | 6.081692 | 458.631309 | 229.980576 | 230.236208 | 11.979058 | 899.175477 |
| **75%** | 1.38800 | 1.38800 | 10.395582 | 573.366747 | 266.574278 | 266.050763 | 16.056631 | 1182.308745 |
| **max** | 1.78400 | 1.78400 | 11.991264 | 654.730219 | 294.239737 | 293.329383 | 17.822302 | 1304.759734 |

Таблица 7. five-summary statistics для сочетания керосин-воздух.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F\_fuel | F\_steam | O2 | CO | NO | NO2 | NO­­x | CO­­2 | SO­2 |
| **count** | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| **mean** | 0.99200 | 0.99200 | 7.233273 | 39.826388 | 47.368964 | 2.742394 | 50.105361 | 10.722833 | 5.347218 |
| **std** | 0.46188 | 0.46188 | 2.704646 | 9.433347 | 22.312363 | 0.478451 | 22.079478 | 2.048165 | 2.094844 |
| **min** | 0.20000 | 0.20000 | 1.857513 | 22.855312 | 23.508647 | 1.916844 | 26.670457 | 8.382153 | 2.934502 |
| **25%** | 0.59600 | 0.59600 | 4.906940 | 31.441446 | 28.607619 | 2.369234 | 31.646136 | 8.894048 | 3.486914 |
| **50%** | 0.99200 | 0.99200 | 8.182716 | 40.000137 | 40.602755 | 2.686694 | 43.390532 | 9.996330 | 4.695978 |
| **75%** | 1.38800 | 1.38800 | 9.651760 | 49.220017 | 59.432063 | 3.175123 | 61.720637 | 12.472772 | 7.129009 |
| **max** | 1.78400 | 1.78400 | 10.327517 | 52.494656 | 98.868625 | 3.480231 | 101.458699 | 14.801686 | 9.597885 |

Таблица 8. five-summary statistics для сочетания керосин-пар.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F\_fuel | F\_steam | O2 | CO | NO | NO2 | NO­­x | CO­­2 | SO­2 |
| **count** | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| **mean** | 0.99200 | 0.99200 | 7.547953 | 44.019278 | 20.499465 | 2.724893 | 23.227495 | 10.426869 | 4.806370 |
| **std** | 0.46188 | 0.46188 | 3.055881 | 19.834181 | 17.198036 | 1.020495 | 17.128364 | 2.299047 | 2.658658 |
| **min** | 0.20000 | 0.20000 | 1.097121 | 14.191735 | 2.818434 | 1.440852 | 5.752742 | 7.968117 | 1.792491 |
| **25%** | 0.59600 | 0.59600 | 5.150211 | 24.426544 | 7.379256 | 1.776850 | 9.969746 | 8.503136 | 2.507365 |
| **50%** | 0.99200 | 0.99200 | 8.844117 | 45.276632 | 14.748060 | 2.509332 | 17.564717 | 9.449069 | 3.879422 |
| **75%** | 1.38800 | 1.38800 | 10.114659 | 62.794588 | 27.005765 | 3.386819 | 30.038772 | 12.207166 | 6.792149 |
| **max** | 1.78400 | 1.78400 | 10.828342 | 73.164397 | 67.257600 | 4.834528 | 69.856680 | 15.333087 | 10.335045 |

Таблица 9. five-summary statistics для сочетания отработанное масло-пар.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F\_fuel | F\_steam | O2 | CO | NO | NO2 | NO­­x | CO­­2 | SO­2 |
| **count** | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| **mean** | 0.99200 | 0.99200 | 3.766710 | 562.014262 | 62.962560 | 12.940645 | 76.006121 | 10.584073 | 113.668501 |
| **std** | 0.46188 | 0.46188 | 1.720883 | 180.739242 | 13.683884 | 6.073395 | 15.022555 | 3.393296 | 16.200912 |
| **min** | 0.20000 | 0.20000 | 1.610373 | 60.935571 | 48.883360 | 4.992585 | 57.696130 | 6.419960 | 85.017342 |
| **25%** | 0.59600 | 0.59600 | 2.176560 | 481.212856 | 51.301857 | 7.307544 | 61.543983 | 7.382732 | 100.497979 |
| **50%** | 0.99200 | 0.99200 | 3.427352 | 609.566035 | 57.153088 | 11.840770 | 71.823291 | 9.527983 | 116.385193 |
| **75%** | 1.38800 | 1.38800 | 5.074379 | 696.187647 | 74.393742 | 18.330662 | 90.346172 | 14.312085 | 128.826700 |
| **max** | 1.78400 | 1.78400 | 6.973130 | 785.525725 | 93.024657 | 23.666318 | 101.577771 | 16.282646 | 134.887662 |

**Визуальный анализ данных**

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

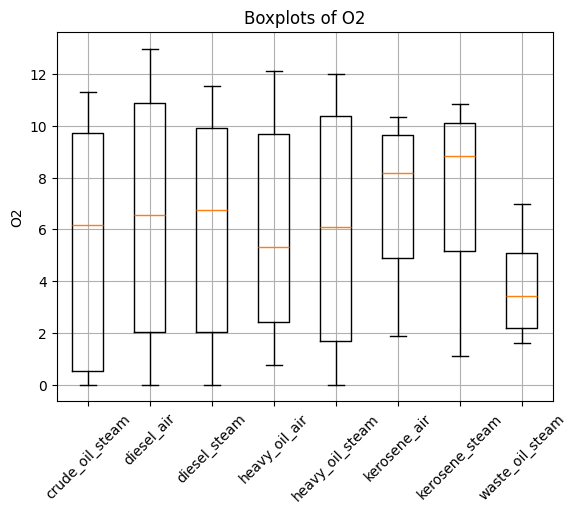


Рис. 2. График boxplot для компонента O2.

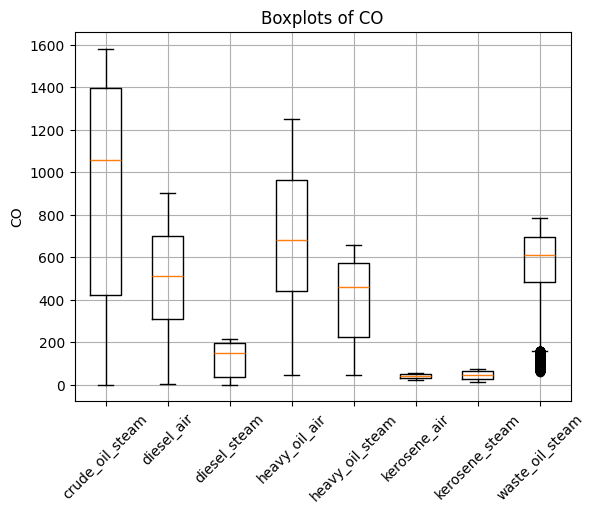


Рис. 3. График boxplot для компонента CO.

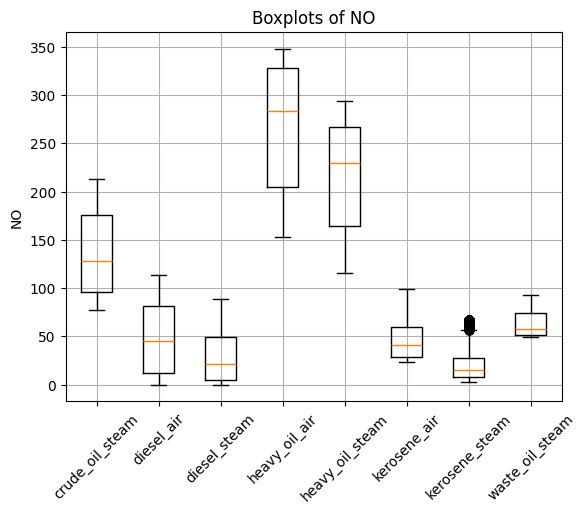


Рис. 4. График boxplot для компонента NO.

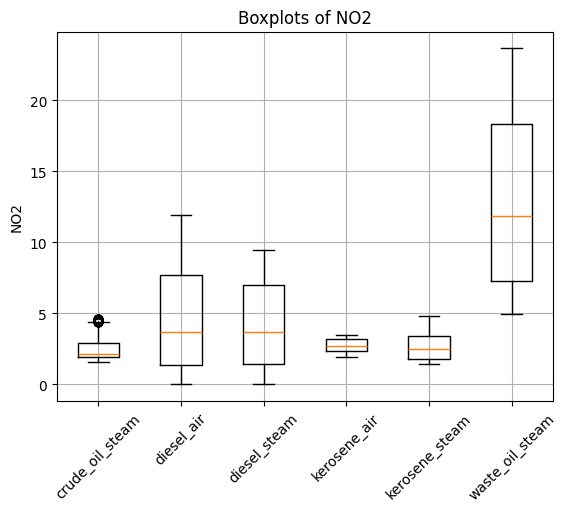


Рис. 5. График boxplot для компонента NO­2.

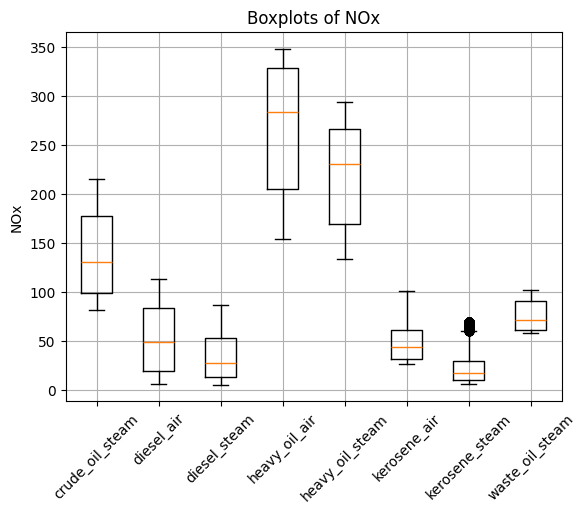


Рис. 6. График boxplot для компонента NO­x.

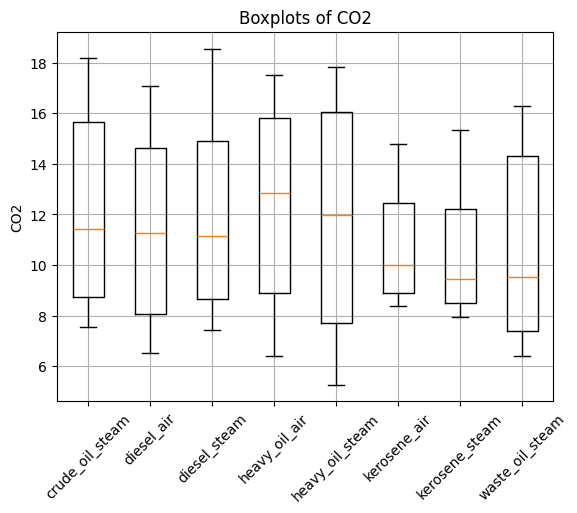


Рис. 7. График boxplot для компонента CO2.

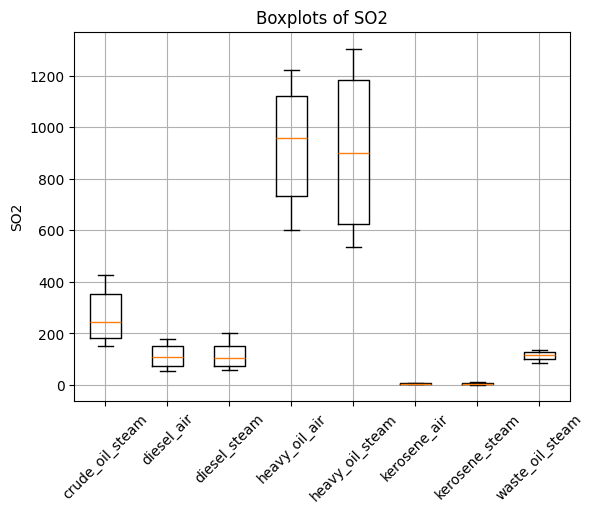


Рис. 8. График boxplot для компонента SO­2.

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, и их можно объяснить режимными особенностями горения.

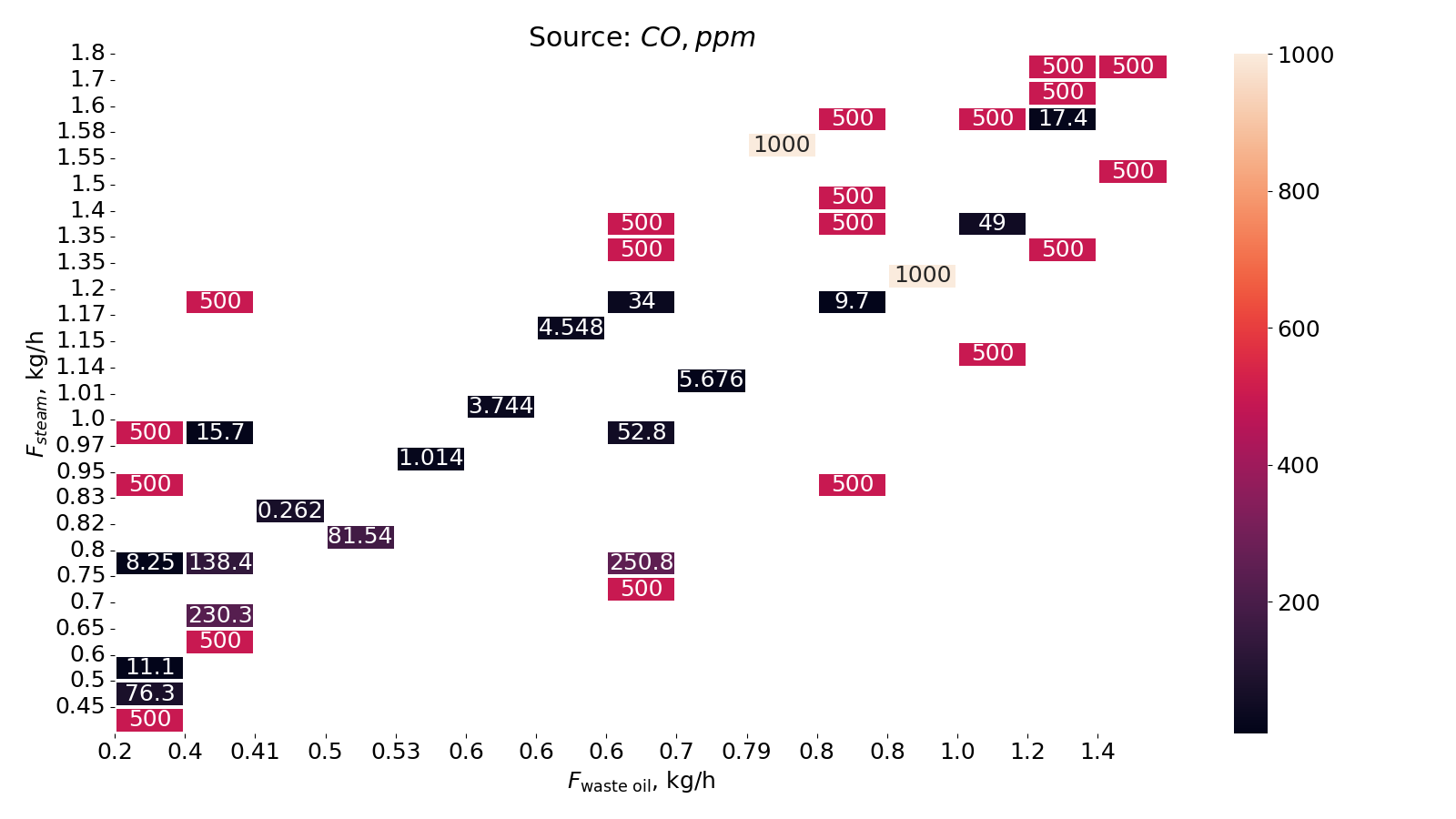


Рис. 9. Опорные точки компонента CO для отработанного масла и вводимого пара.

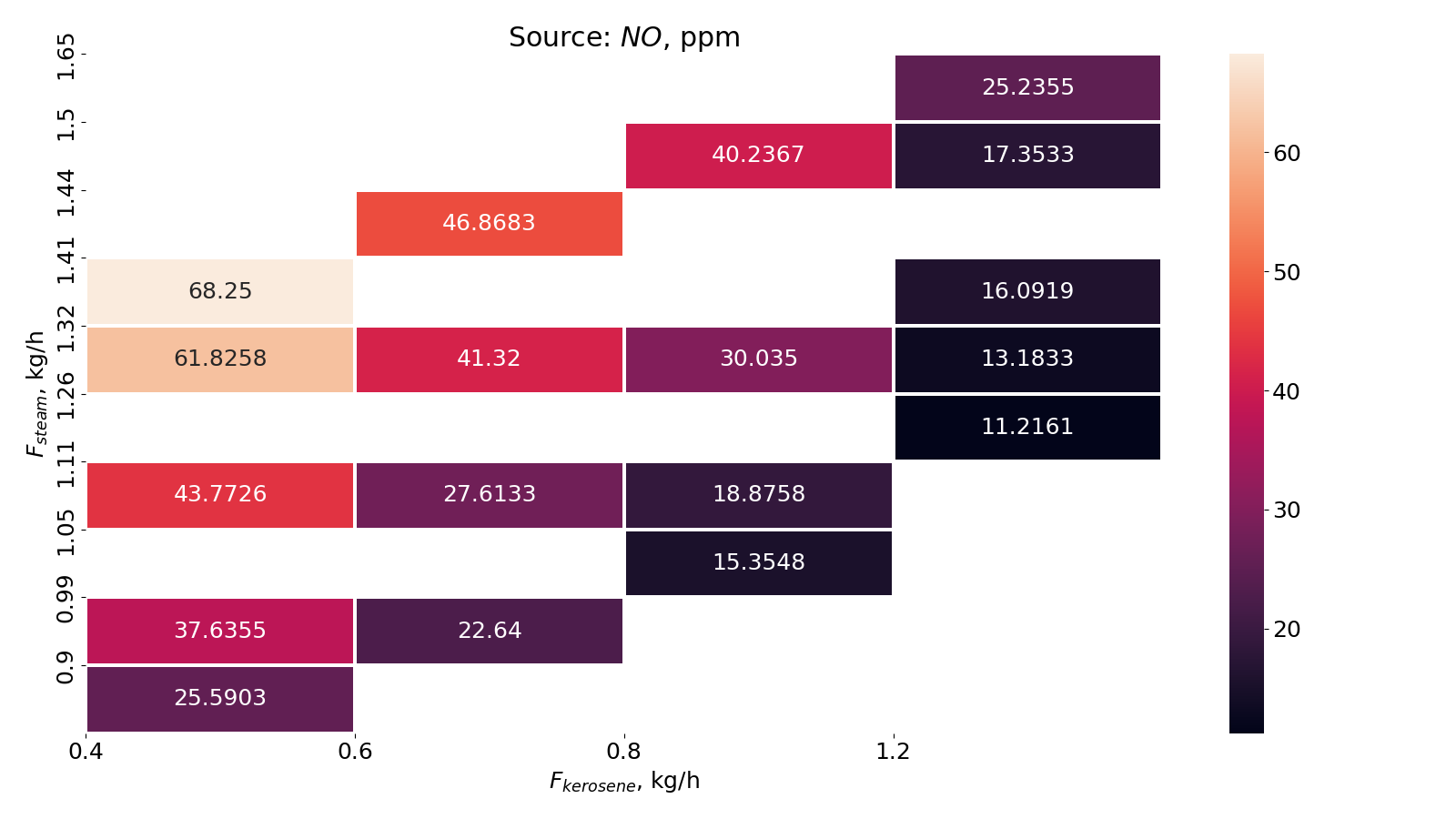


Рис. 10. Опорные точки компонента NO для керосина и вводимого пара.

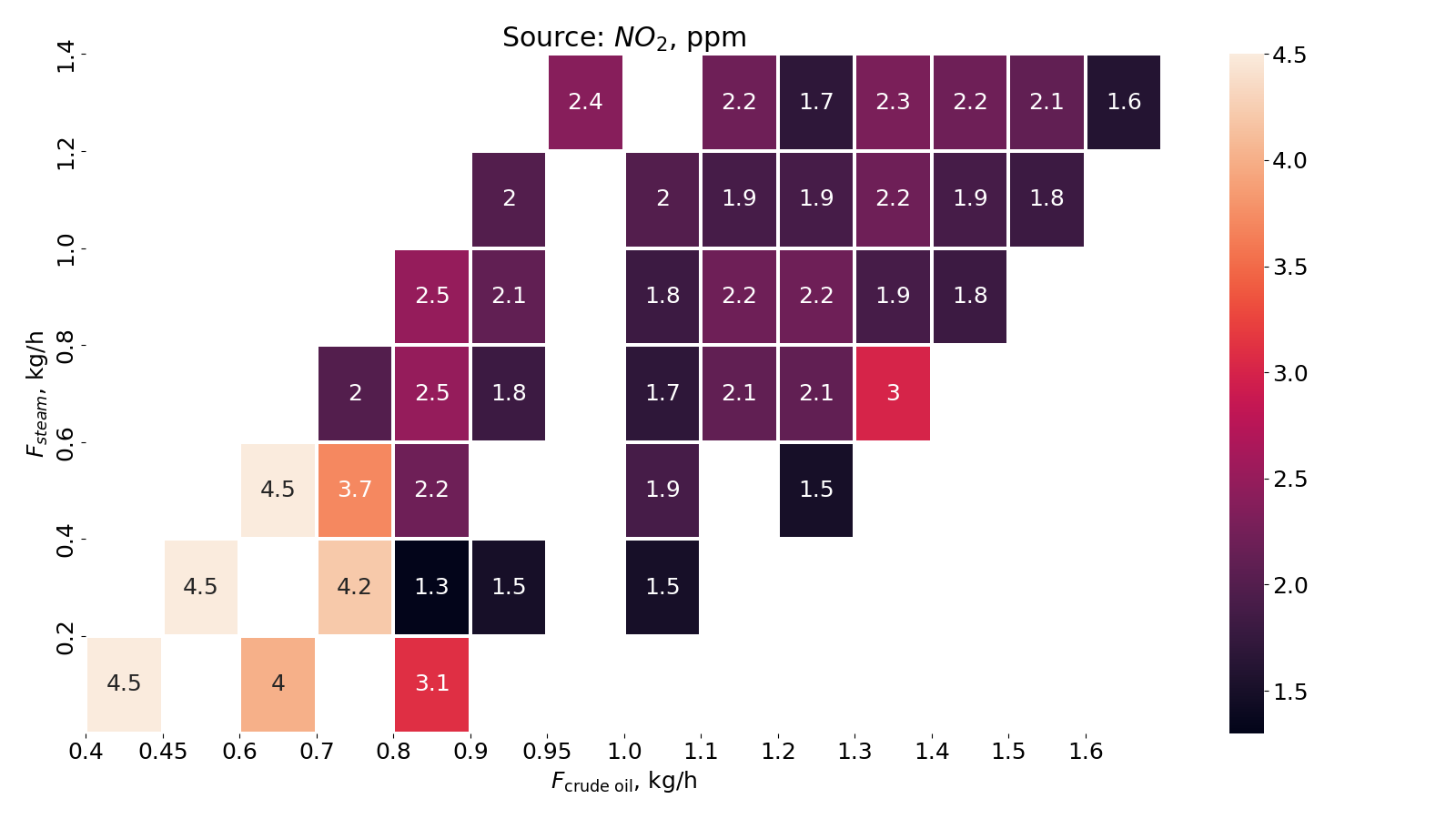


Рис. 11. Опорные точки компонента NO2 для керосина и вводимого пара.

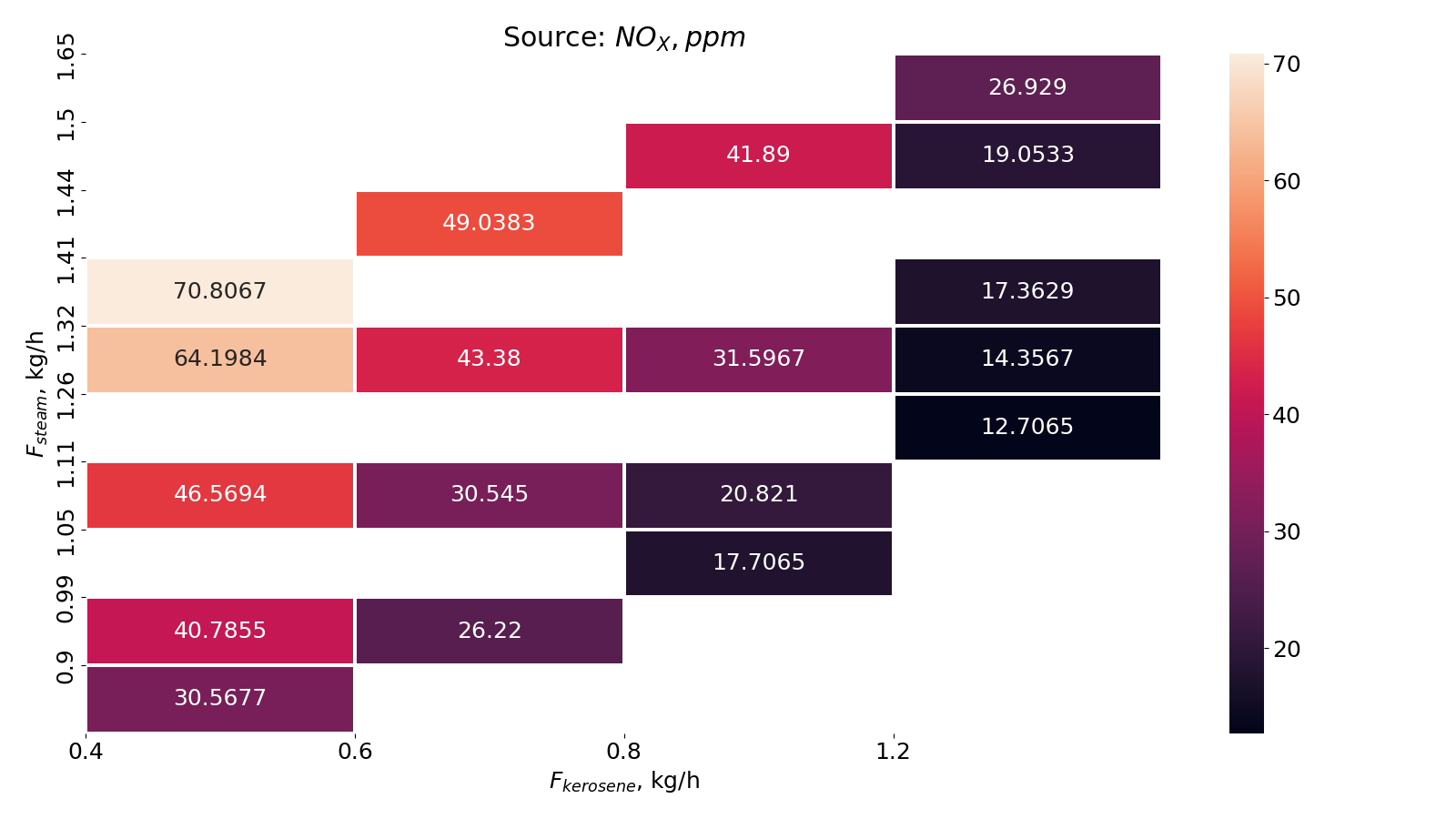


Рис. 12. Опорные точки компонента NOx для керосина и вводимого пара.

Далее рассмотри 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.

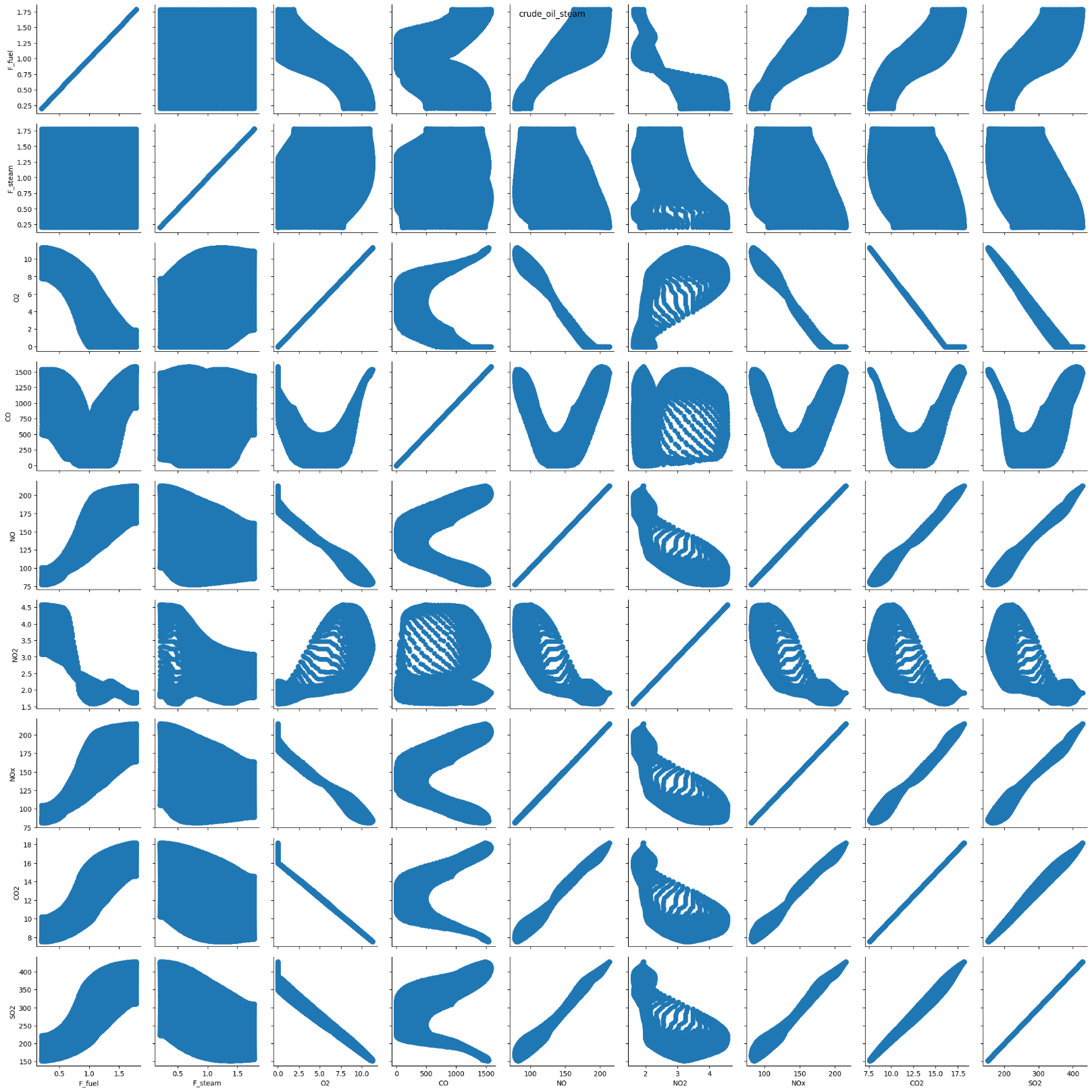


Рис. 13. График парных сеток для сырой нефти и вводимого в пламя пара.

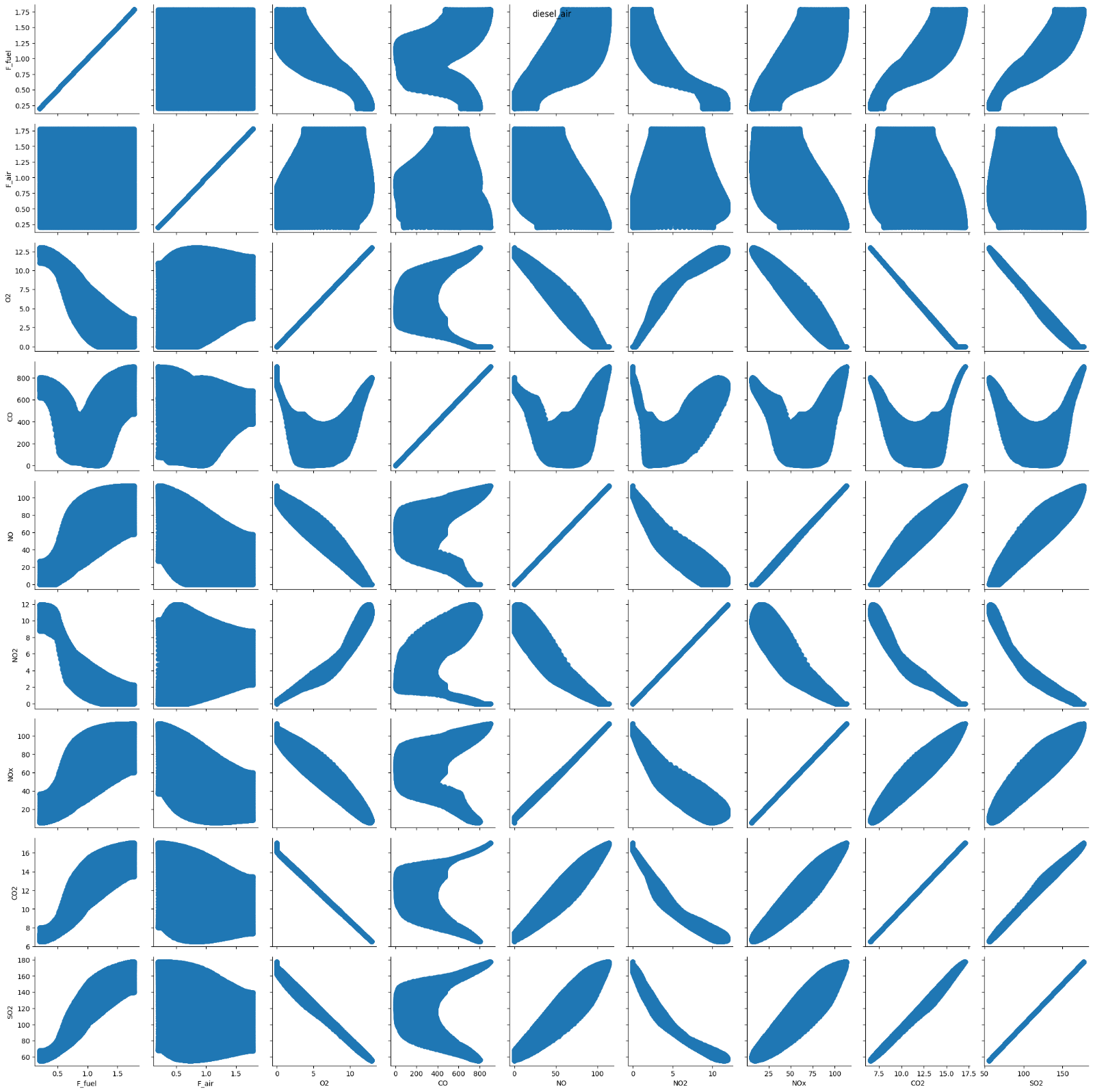


Рис. 14. График парных сеток для дизельного топлива и вводимого в пламя воздуха.

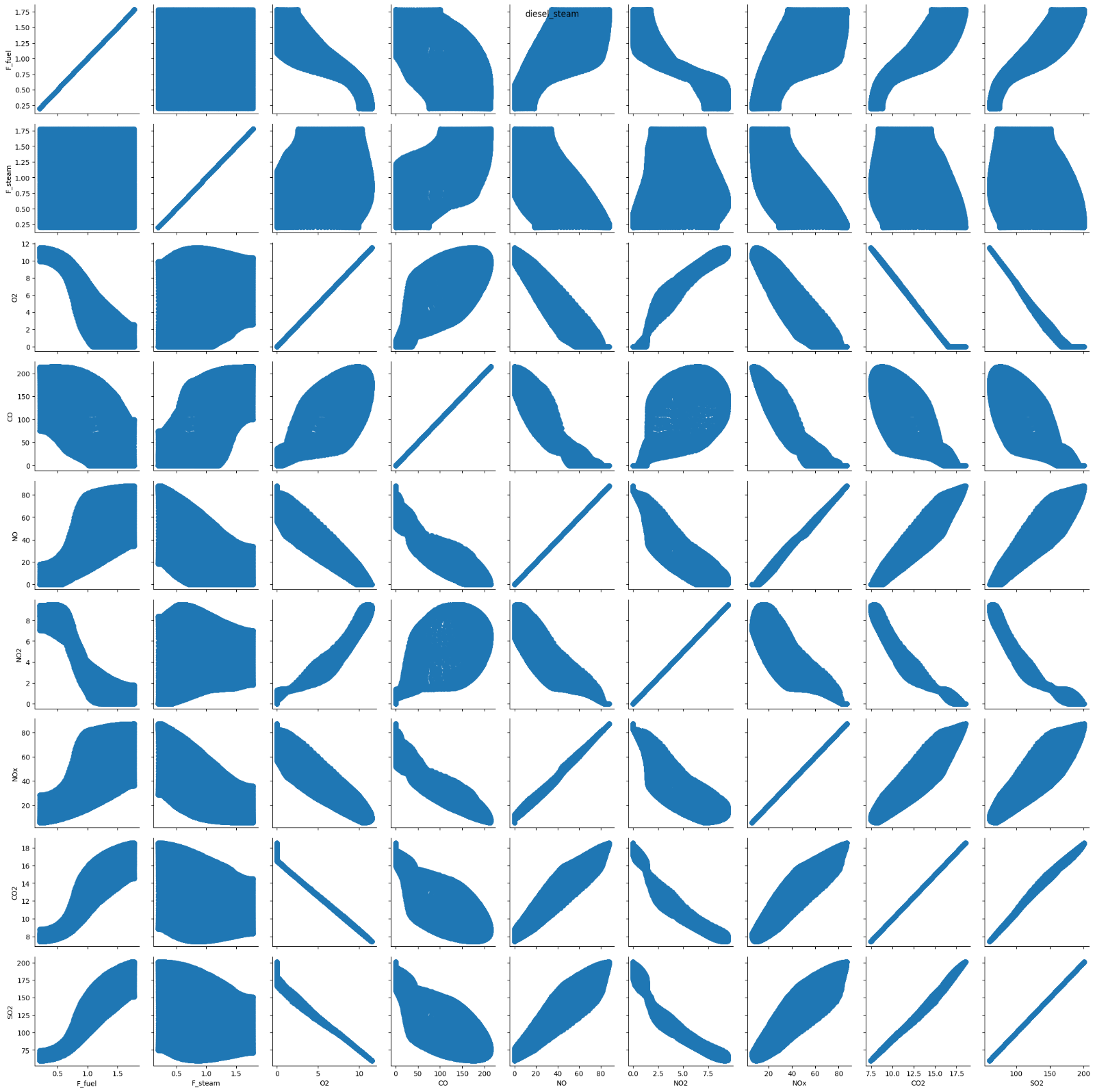


Рис. 15. График парных сеток для дизельного топлива и вводимого в пламя пара.

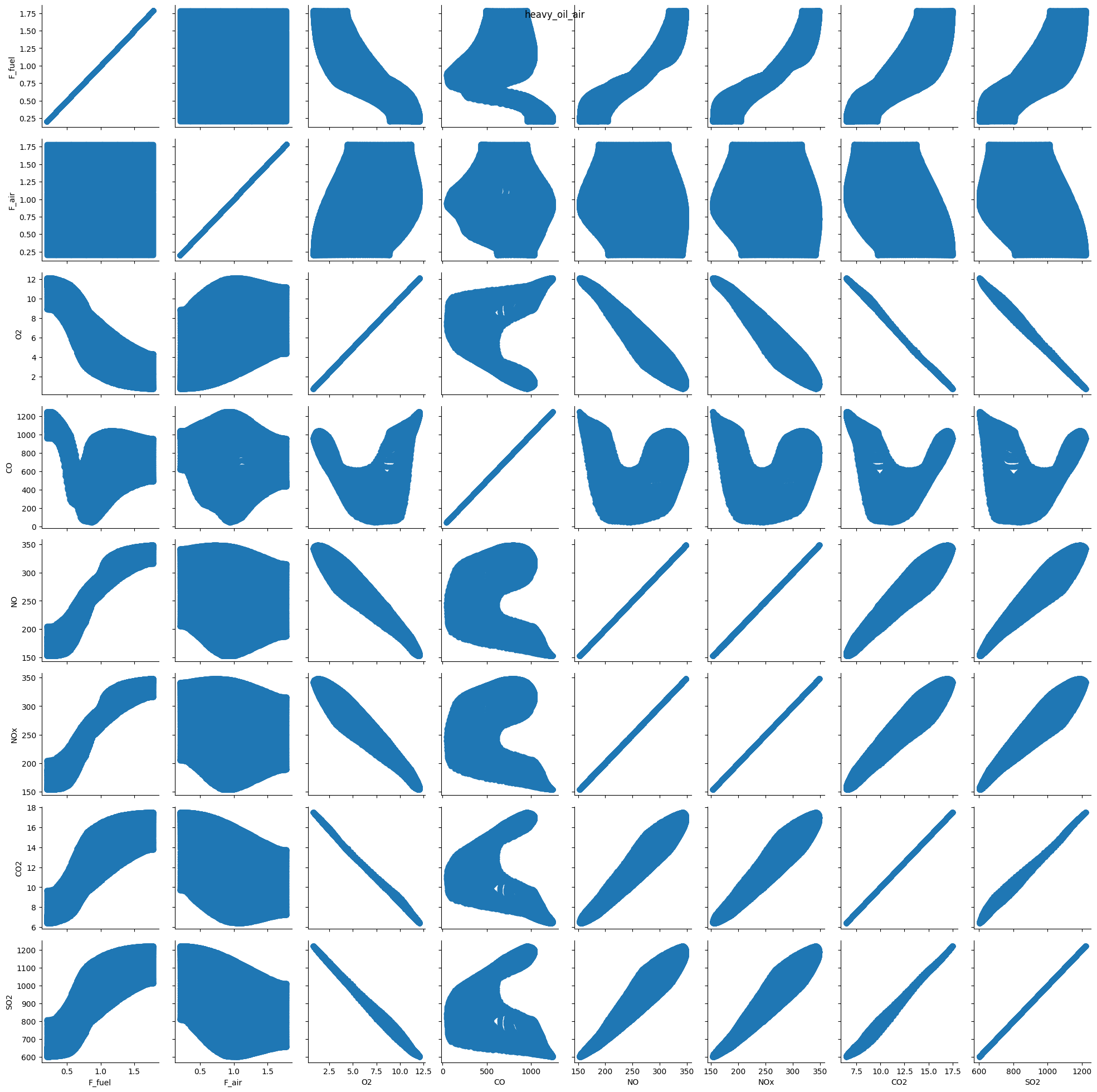


Рис. 16. График парных сеток для мазута и вводимого в пламя воздуха.

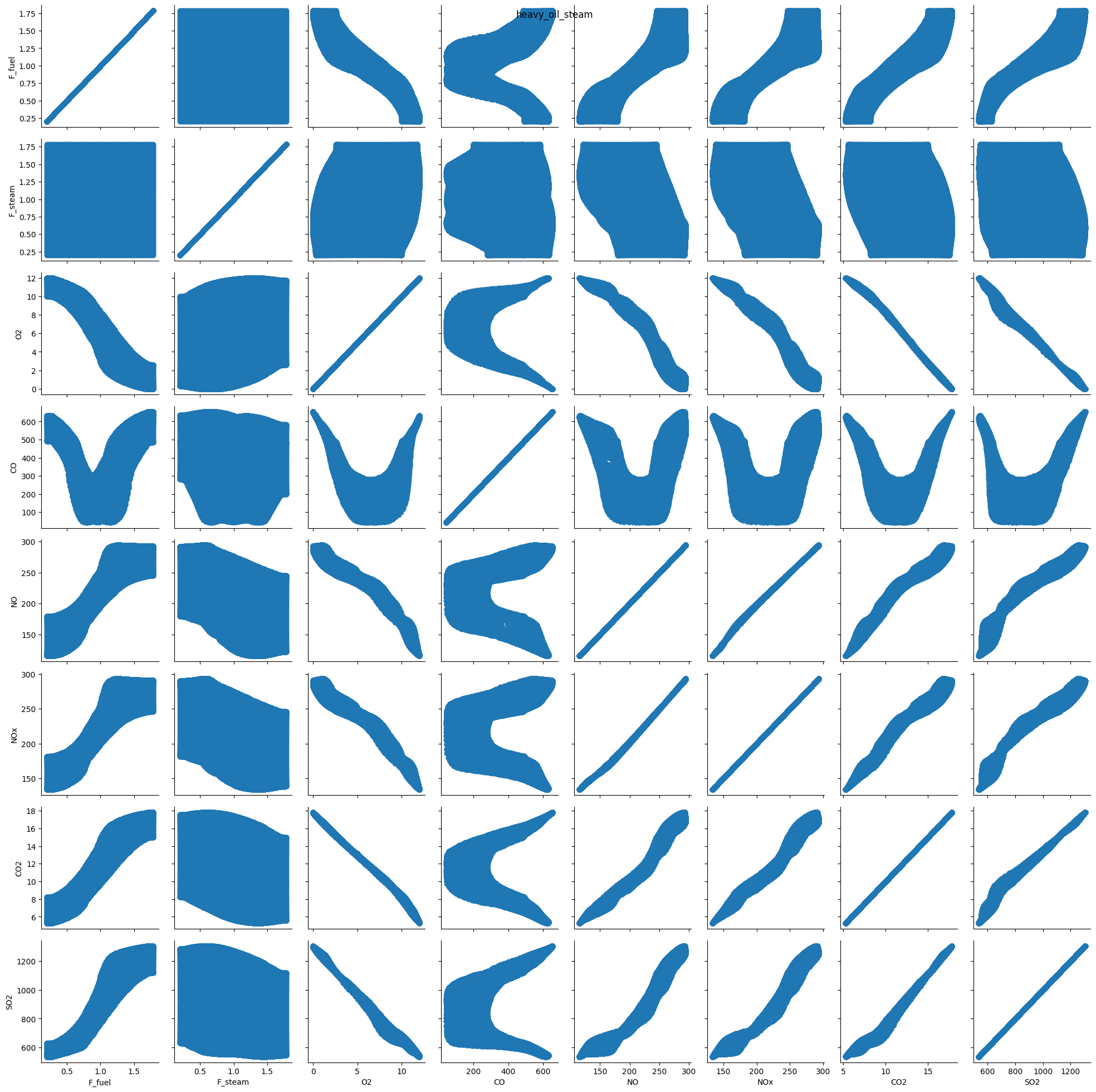


Рис. 17. График парных сеток для мазута и вводимого в пламя пара.

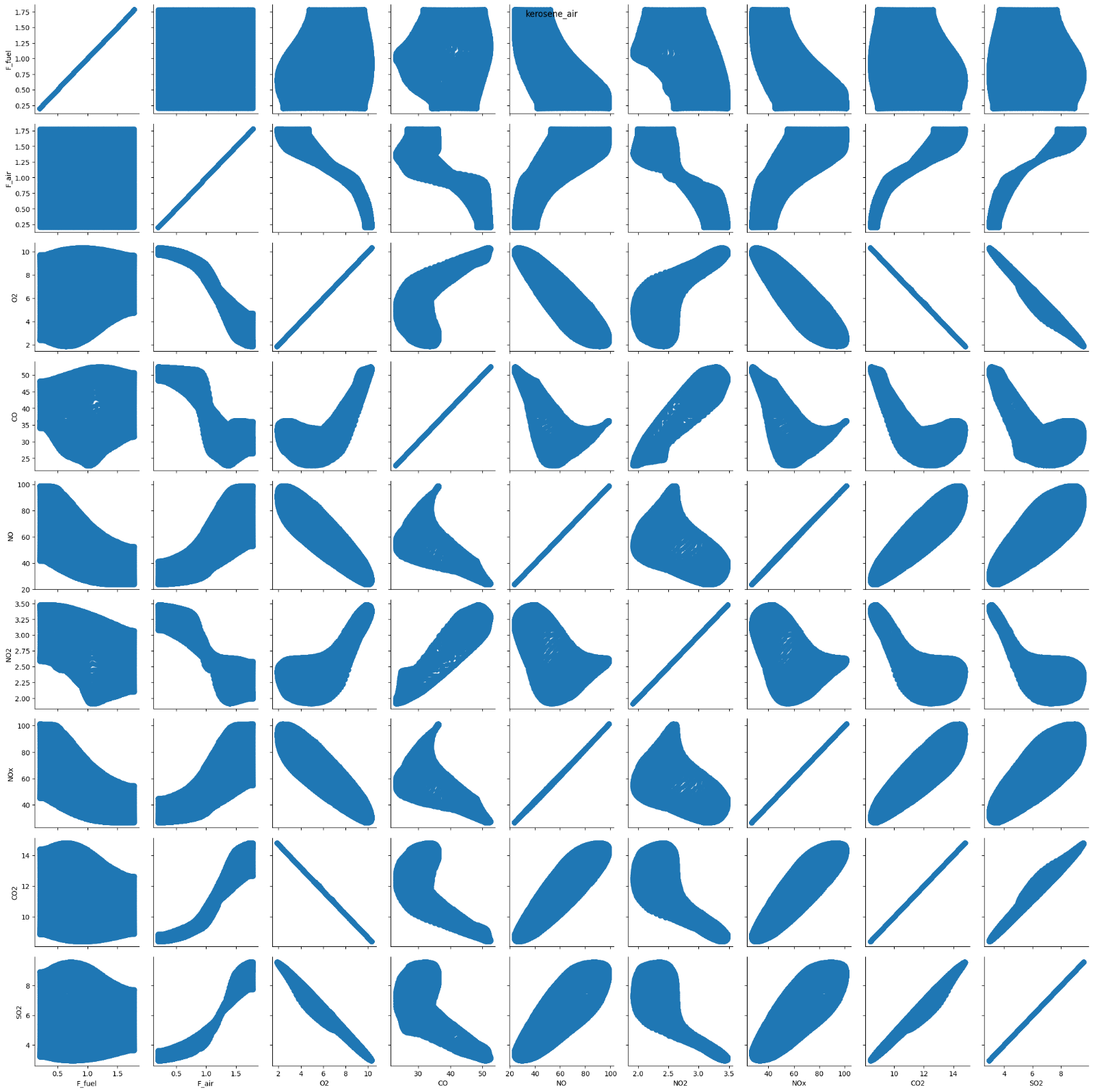


Рис. 18. График парных сеток для керосина и вводимого в пламя воздуха.

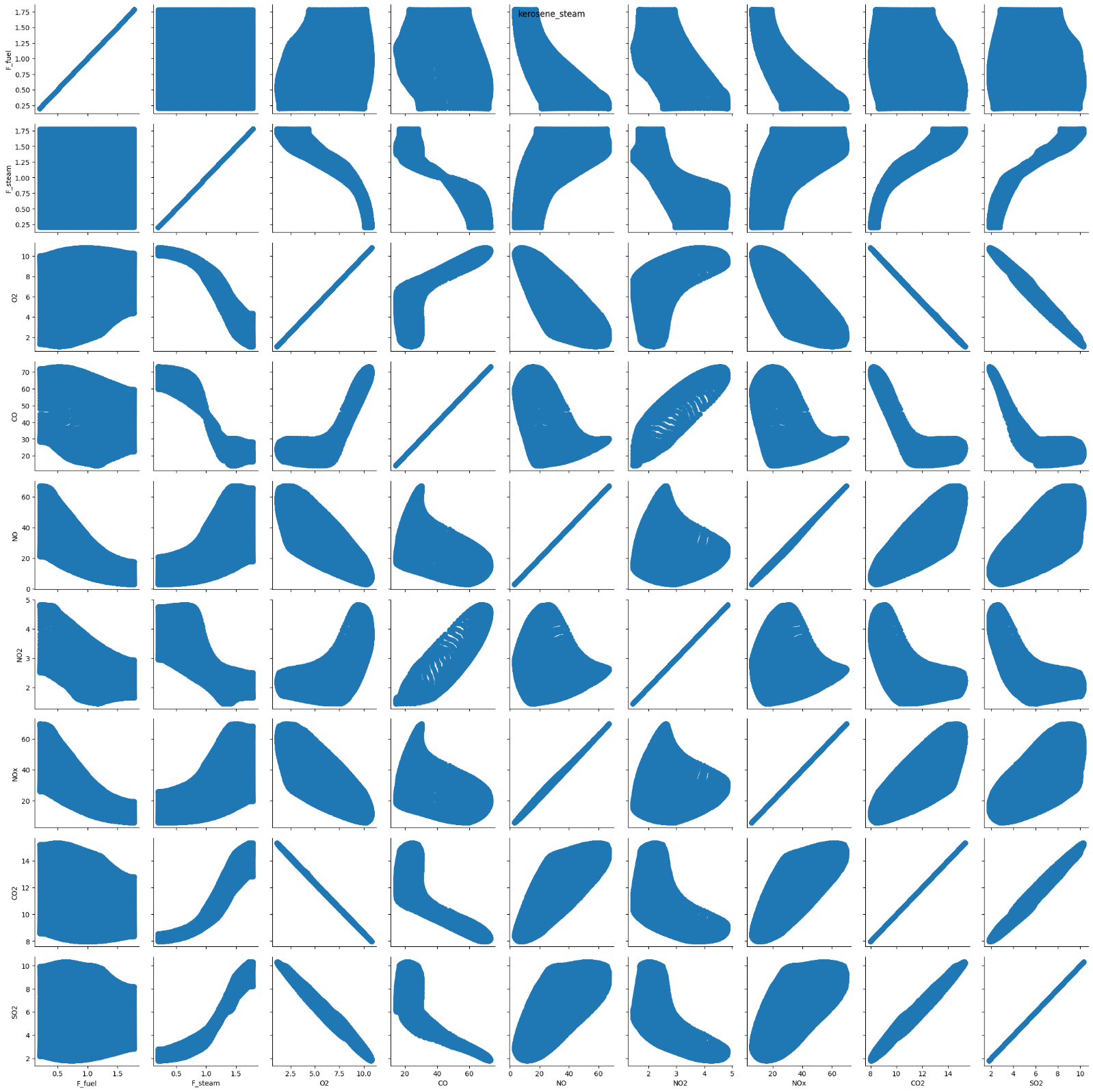


Рис. 19. График парных сеток для керосина и вводимого в пламя пара.

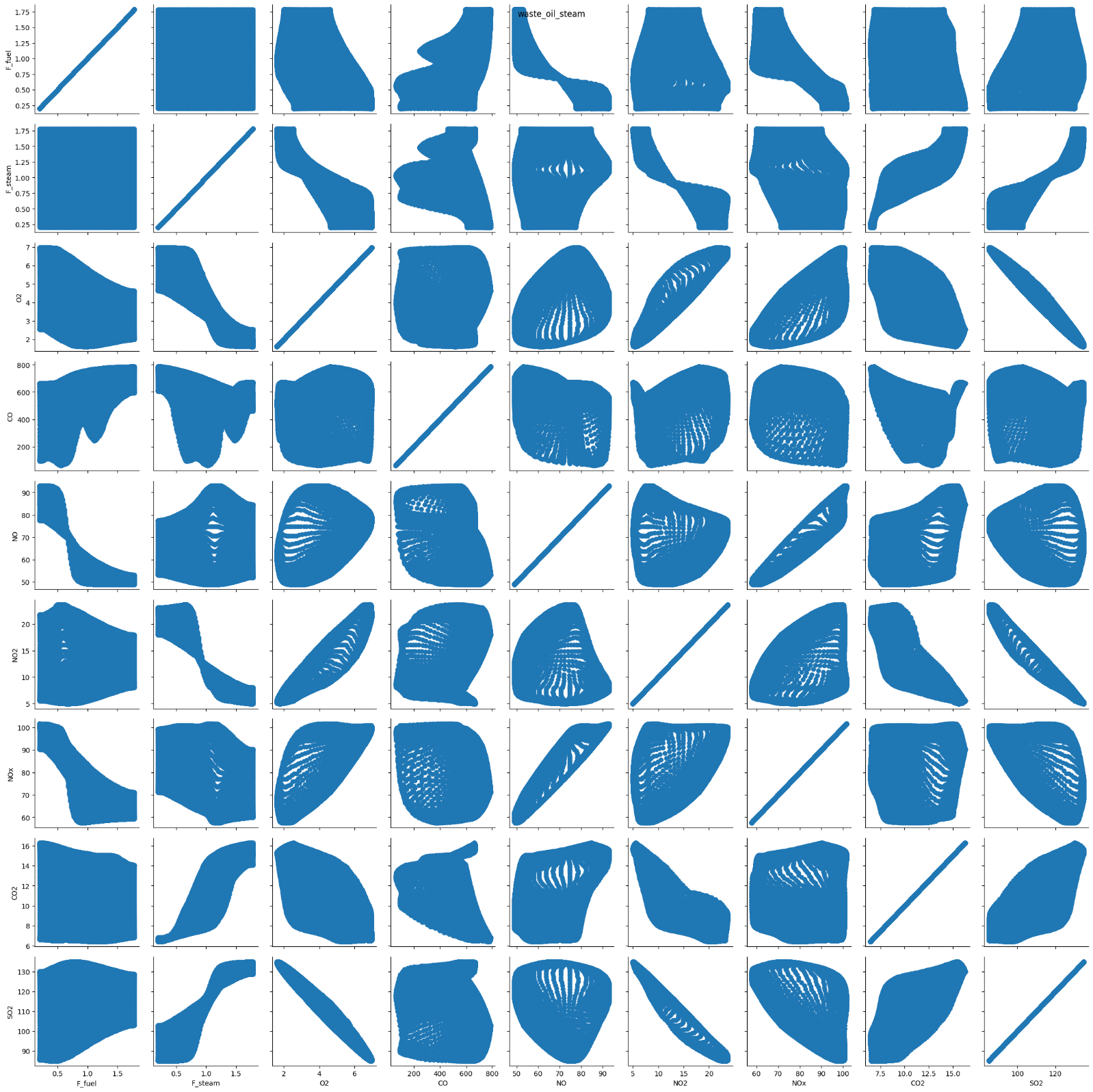


Рис. 20. График парных сеток для отработанного масла и вводимого в пламя пара.

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.

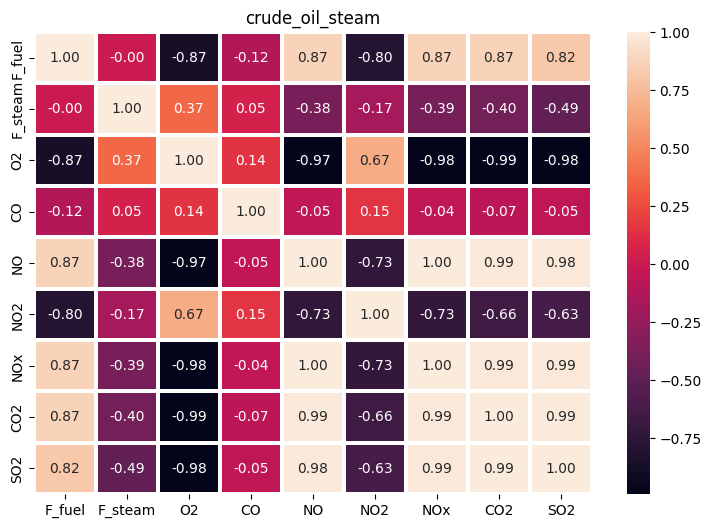


Рис. 21. Корреляционная матрица для сырой нефти и вводимого в пламя пара.

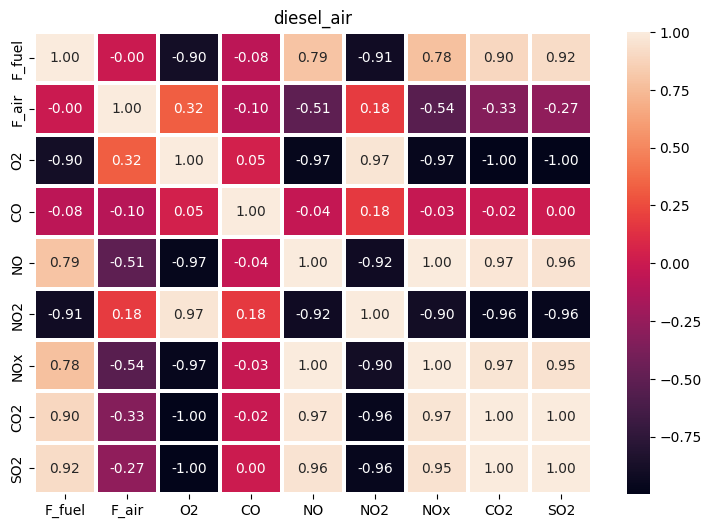


Рис. 22. Корреляционная матрица для дизельного топлива и вводимого в пламя воздуха.

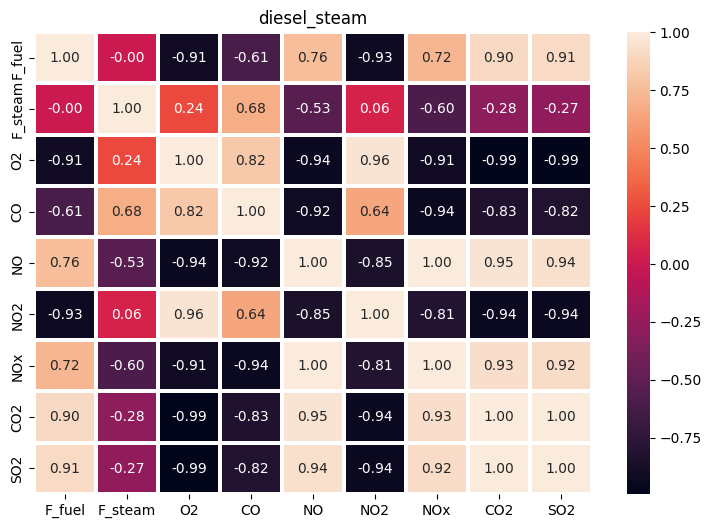


Рис. 23. Корреляционная матрица для дизельного топлива и вводимого в пламя пара.

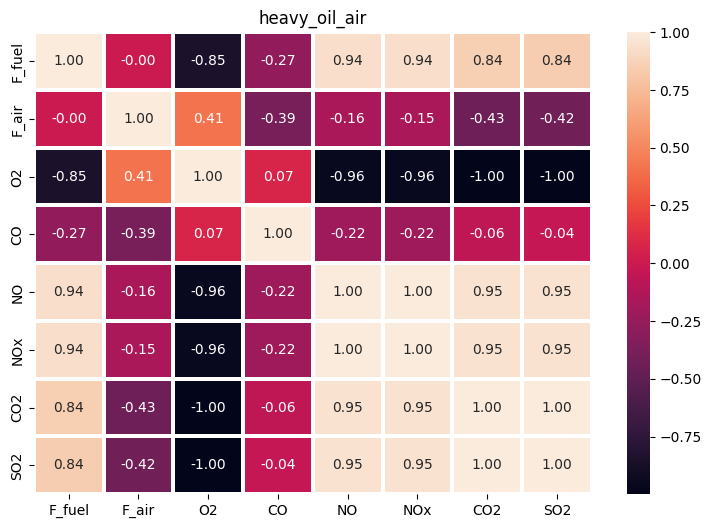


Рис. 24. Корреляционная матрица для мазута и вводимого в пламя воздуха.

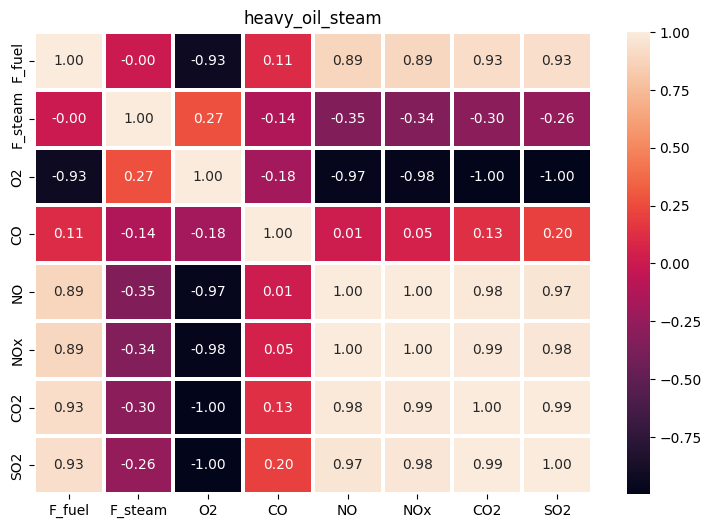


Рис. 25. Корреляционная матрица для мазута и вводимого в пламя пара.

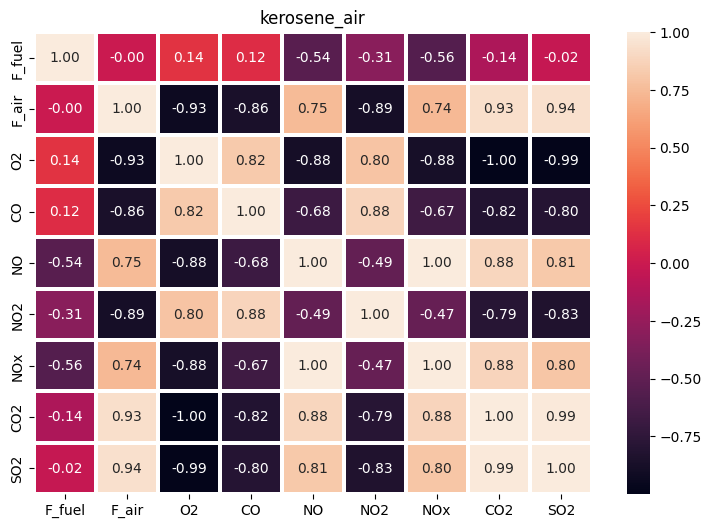
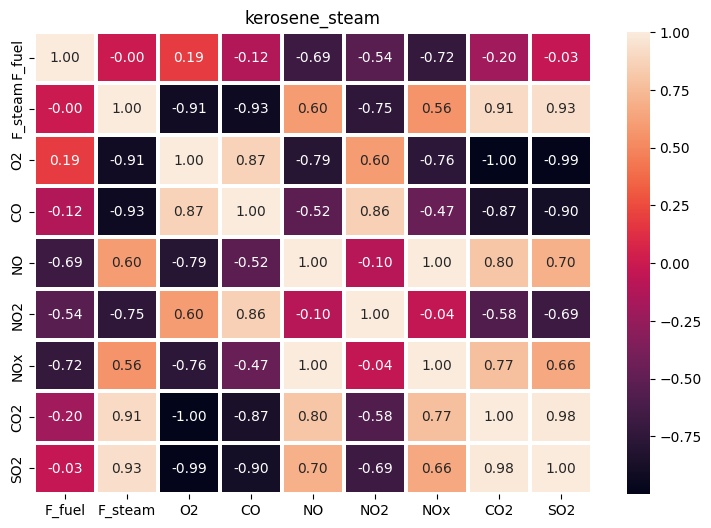
Рис. 26. Корреляционная матрица для мазута и вводимого в пламя воздуха.

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

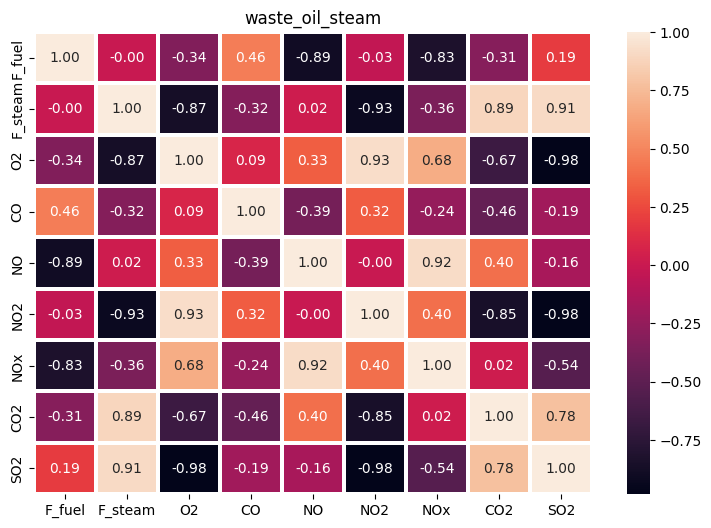


Рис. 28. Корреляционная матрица для отработанного масла и вводимого в пламя пара.