Содержание

1 Описание данных и постановка задачи

Имеются данные о находках на поселенческом памятнике. Каждая строка данных состоит из трех координат, задающих место обнаружения предмета, названия археологического слоя, в котором он залегал, названия категории предмета и типа его сырья. Кроме того, имеется бинарный признак обожженности находки. Основная задача состоит в нахождении источников концентрации термически обработанных находок.
Математически задача сводится к ряду однотипных задач следующего вида. Имеются два набора трехмерных данных (точек в пространстве). Требуется найти области, в которых концентрация точек одного из наборов статистически значимо отличается от концентрации другого набора. Здесь возникает ряд сложностей с тем, что данных может быть относительно небольшое количество, а алгоритмы могут переобучаться и находить незначимые (случайные) отклонения.

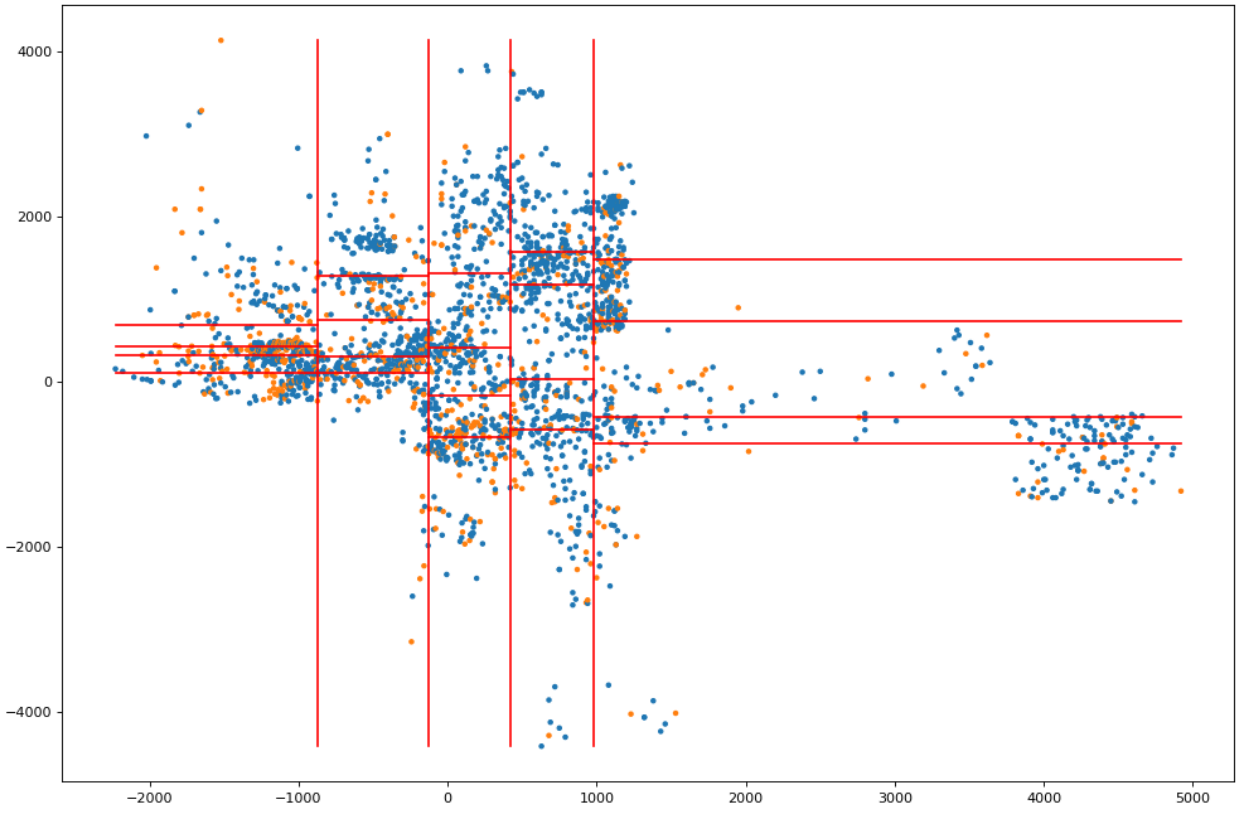
2 Используемые методы

2.1 Проверка однородности

Метод заключается в проверке однородности с помощью критерия хи-квадрат. В чем он состоит?

- Первый вариант: к отдельным слоям/участкам применяется двумерный критерий (считая, что берется тонкий равномерный по оси z слой).

  Суть применения двумерного критерия заключается в разбиении всей плоскости на прямоугольники так, чтобы в каждый попадало примерно одинаковое число находок. Способ разбиения на эти прямоугольники является неоднозначным и может настраиваться как гиперпараметр. Пример одного из возможных разбиений частиц в палево-желтом слое (при анализе было использовано другое, об этом подробнее в результатах):

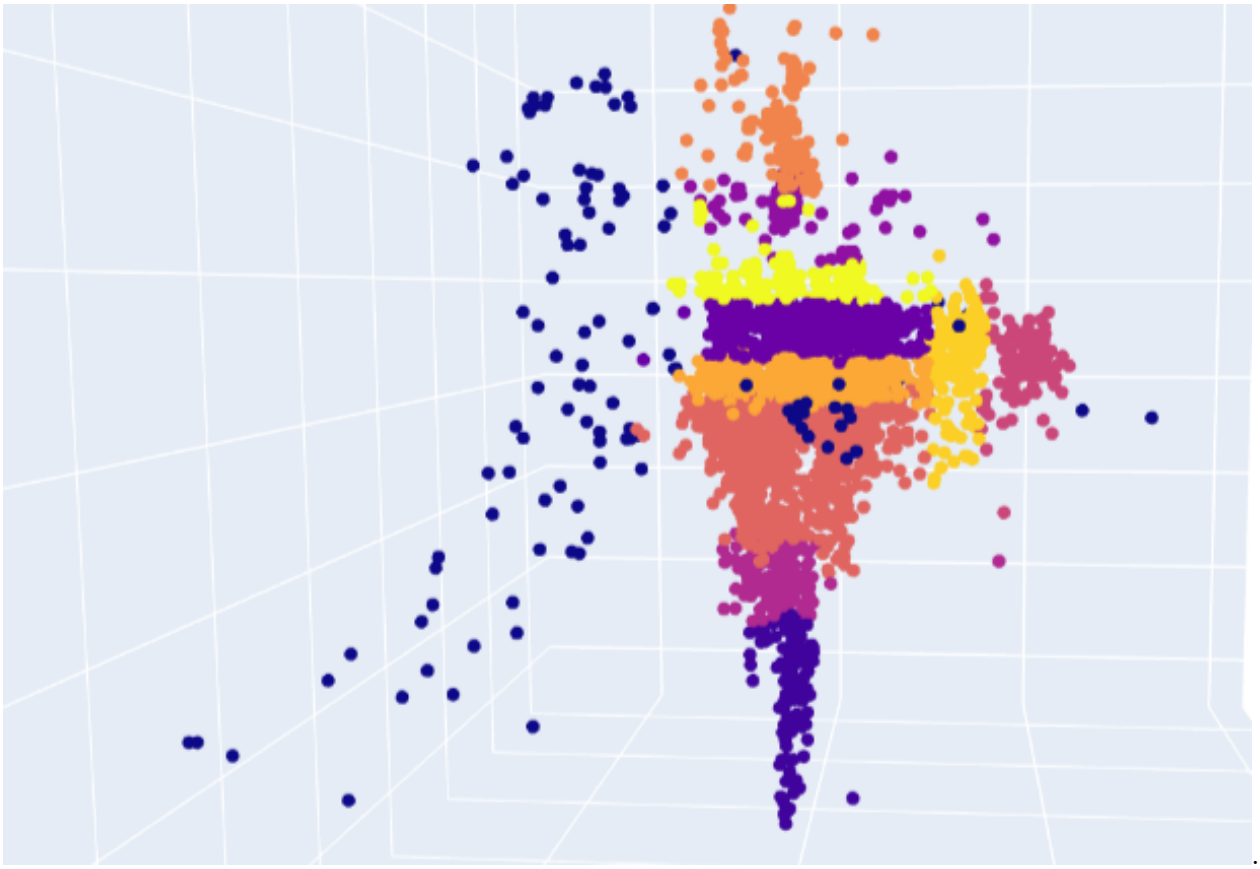


Далее внутри каждого прямоугольника вычисляется процент обожженных находок и к полученным значениям применяется критерий однородности хи-квадрат.

- Второй вариант: ко всем данным применяется трехмерный критерий хи-квадрат.

  В трехмерном варианте нужно разбивать уже с учетом оси z, соответственно вместо прямоугольников получаем некие параллелепипеды. В данном случае деление было в основном ручное, с учетом особенностей "рельефа" , чтобы выделить в отдельные ячейки значимые скопления. А далее все аналогично двумерному случаю: считается число обожженных и не обожженных частиц и проверяется однородность с помощью критерия хи-квадрат.

  Разбиение имело следующий вид:



В трехмерном случае границы ячеек подбирались так, что в каждую попало примерно по 160 находок.

- Кроме того, был проведен анализ статистической значимости неоднородных особенностей. Для этого находки случайным образом много раз делились на обожженные и необожженные (в той же пропорции, что и в исходной выборке). После этого для каждой раскраски вычислялась статистика критерия хи-квадрат. И в конце считалось, в скольких раскрасках она получилась больше, чем статистика, посчитанная по исходным данным, а в скольких меньше. На основе этого получали новое p-value. Этот способ позволяет лишь доказать наличие однородности.

- Был также применен метод, позволяющий обнаружить в каких именно областях наблюдаются неоднородности. Для этого опять много раз находки были случайно разделены на обожженные и не обожженные. После этого каждый раз запоминали какое максимальное число обожженных и не обожженных находок попало в одну ячейку при данной раскраске. Эти числа были упорядочены, и отброшены 5% самых больших из них (так как эти значения могли случайно оказаться очень большими). Наибольшее значение в оставшемся списке будем считать ориентировочным максимумом случайного числа обожженных и не обожженных частиц. После этого все ячейки изначальной раскраски, в которые попало больше, чем

пороговое значение обожженных частиц раскрасим в красный, а не обожженных в синий. Одновременно оба случая реализоваться в одной ячейке не могут, так как суммарное число находок в каждой ячейке одинаковое. Красные области - потенциальные очаги обожженных частиц.

2.2 Ядерные оценки плотностей

Предложенный метод уязвим к выбору дискретных областей и их размеров, поэтому рассмотрим альтернативный подход. В нём мы оценим "плотность" находок в окрестности каждой точки, используя стандартный статистический метод ядерного оценивания. В результате в каждой точке мы найдем плотность обожженных и необожженных находок, а затем посмотрим, насколько эти значения отличаются, и промаркируем те области, где отличия вызваны разницей концентраций находок, а не просто их случайной флюктуацией. Для этого применяется метод бутстрэп: производится многократная генерация выборок на базе имеющихся данных, построение по ним ядерных оценок плотностей и их разностей, а также вычисление отклонения от изначально полученного результата. В результате удается построить равномерный доверительный интервал для разности плотностей и выделить статистически значимые различия в распределениях данных. Обоснование корректности этой процедуры приведено в [1].

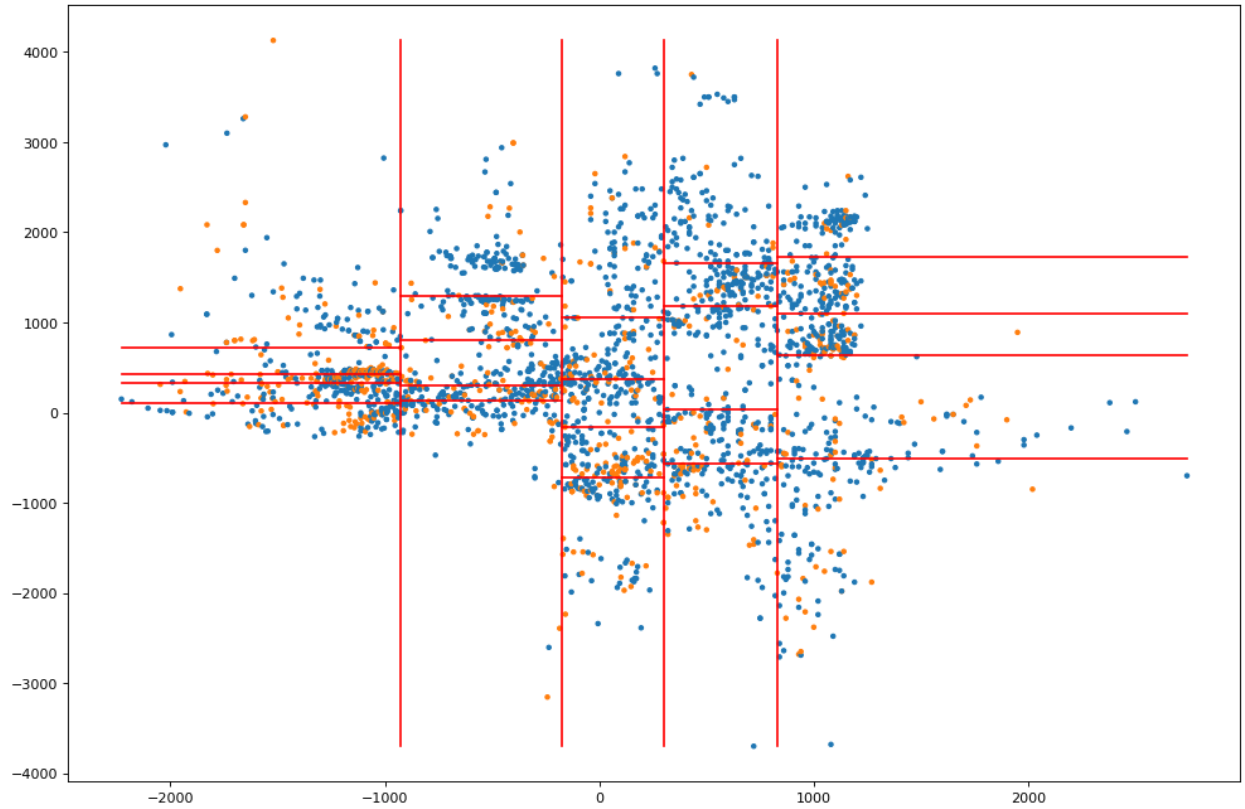
3 Применение методов

3.1 Проверка однородности

- **Обработка палево-желтого слоя.** Описанным выше методом были сначала обработаны находки, обнаруженные в палево-желтом слое. Их общий вид представлен на рисунке, где оранжевый отвечает за обожженные находки.

.

Видно, что все находки в этом слое можно разбить на две части: основная - та, которая выделена прямоугольником, и небольшое скопление, обозначенное кругом, остальные для упрощения работы исключим из рассмотрения. Далее находки основной части были разбиты на ячейки так, что в каждой примерно одинаковое число находок. Получилось следующее:



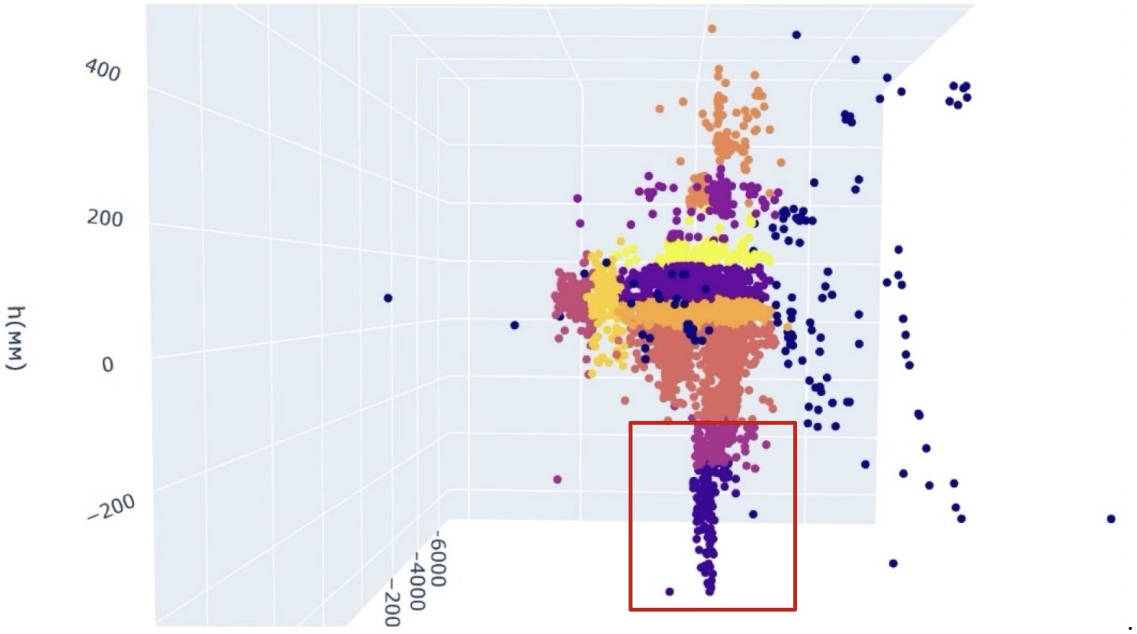
После анализа статистики критерия хи-квадрат для данного разбиения можно сделать вывод о том, что соотношение обожженных и не обожженных частиц в нем неоднородно. В верхнем правом углу основной (в проекции на плоскость OXY) меньше всего обожженных находок, и в целом справа их меньше, чем слева, при том что общее число частиц в ячейках одинаково. В отдельном небольшом скоплении также сравнительно мало обожженных находок. Можно прежположить, что источник обожженных частиц находился ближе к левому нижнему углу.

- **Проверка статистической значимости.** Анализ показал, что при случайных раскрасках значения статистики получаются в разы меньше, чем на расссматриваемых данных. За 1000 перекрасок статистика не превысила 40, а исходная статистика равнялась 214. Это говорит о том, что **в палево-желтом слое находки, действительно, распределены неравномерно и этот результат является статистически значимым.**

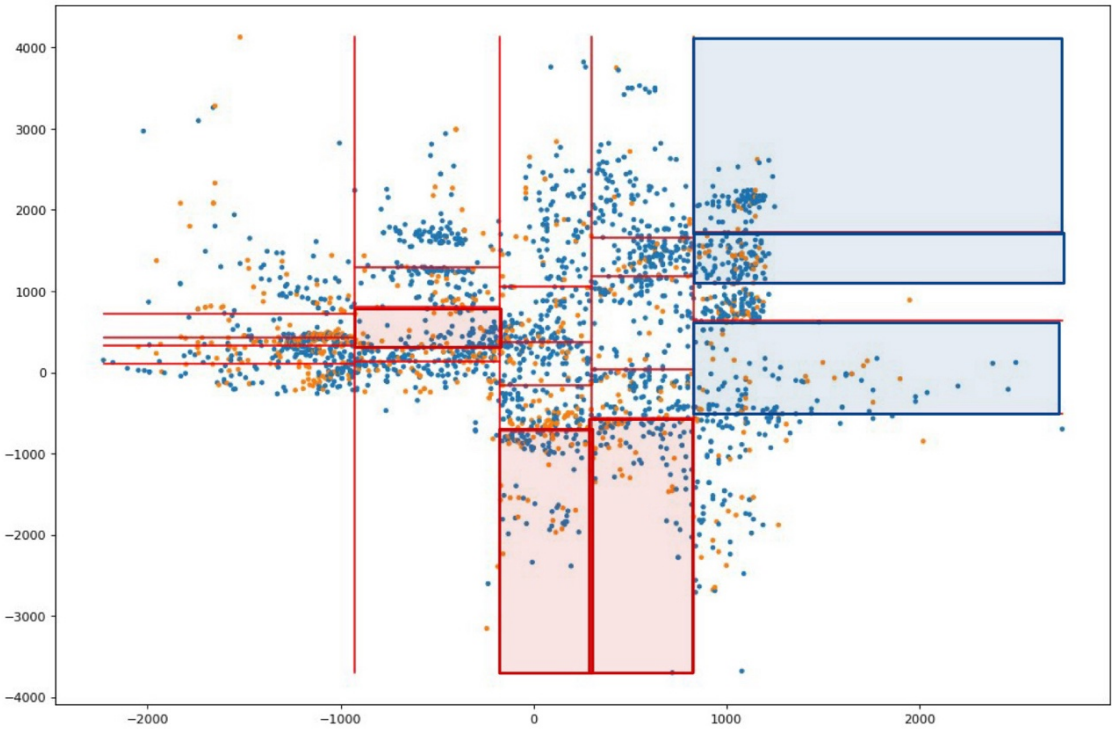
- **Другие слои.** Кроме того, была проведена двумерная обработка находок из слоя желтого песка и границы желтого песка с палево-желтым. В низ было довольно мало находок и явные особенности обнаружить не удалось.

- **Обработка ямы**. Также двумерным методом была обследована "визуальная яма". Было проведено ее исследование в проекции на плоскость OXY, OXZ, OYZ. В проекции на каждую из осей яма оказалась однородной, поэтому далее участвовала в общем трехмерном исследовании, без учета особенности рельефа. Скопление то-

чек, которое мы называем "ямой" (обведено в прямоугольник на иллюстрации):



- **Трехмерное исследование**. Применение трехмерного критерия хи-квадрат показало, что обожженные находки распределены очень неравномерно, но выделить конкретные области особенностей на данном этапе исследования не удалось, также как и проверить статистическую значимость неоднородности.
- **Выявление статистически значимых участков неоднородности в палево-желтом слое.** Результаты представлены на картинке. Здесь красным выделены участки, на которых число обожженных находок больше порогового значения (определен в описании метода), а синим выделены участки, где число не обожженных находок больше порогового значения.



3.2 Ядерные оценки плотностей

Для выявления различий в расположении термически обработанных и необработанных находок, координаты были переведены в метрическую шкалу (сжаты в 1000 раз), а уровень доверия всюду был установлен на 0.95. Это означает, что указанные нами отличия с вероятностью 95% возникли не в результате случайности, то есть они обусловлены реальным различием концентраций находок.
Рассмотрены плоские проекции археологического памятника (вид сверху) по слоям, по категории находок и по типу сырья. На первых двух графиках изображаются полученные оценки плотностей обожженных и необожженных находок в конкретном слое. Участки повышенной концентрации тех или иных находок соответствуют красному цвету, пониженной – синему. Затем приводится график разности этих плотностей с указанием статистической погрешности.
Также описаный метод был применен в трехмерном случае. В нем мы изображаем рассматриваемую группу находок зеленым цветом, выделяя находки, попавшие в значимые области повышенной концентрации обожженного материала красным цветом, а пониженной – синим.

3.2.1 По слоям

Количество находок позволило провести анализ в слоях палево-желтого песка и желтого песка.
При плоском анализе в палево-желтом песке обнаружен участок концентрации терми-

чески обработанных находок, на нижнем графике рис. 1 он выделяется темно-красным цветом. Пространственный анализ (см. рис. 2) подтвердил этот результат, а также позволил обнаружить менее крупное место концентрации обожженного материала. Отметим, что более половины находок, отмеченных красным цветом на рис. 2, являются термически обработанными.

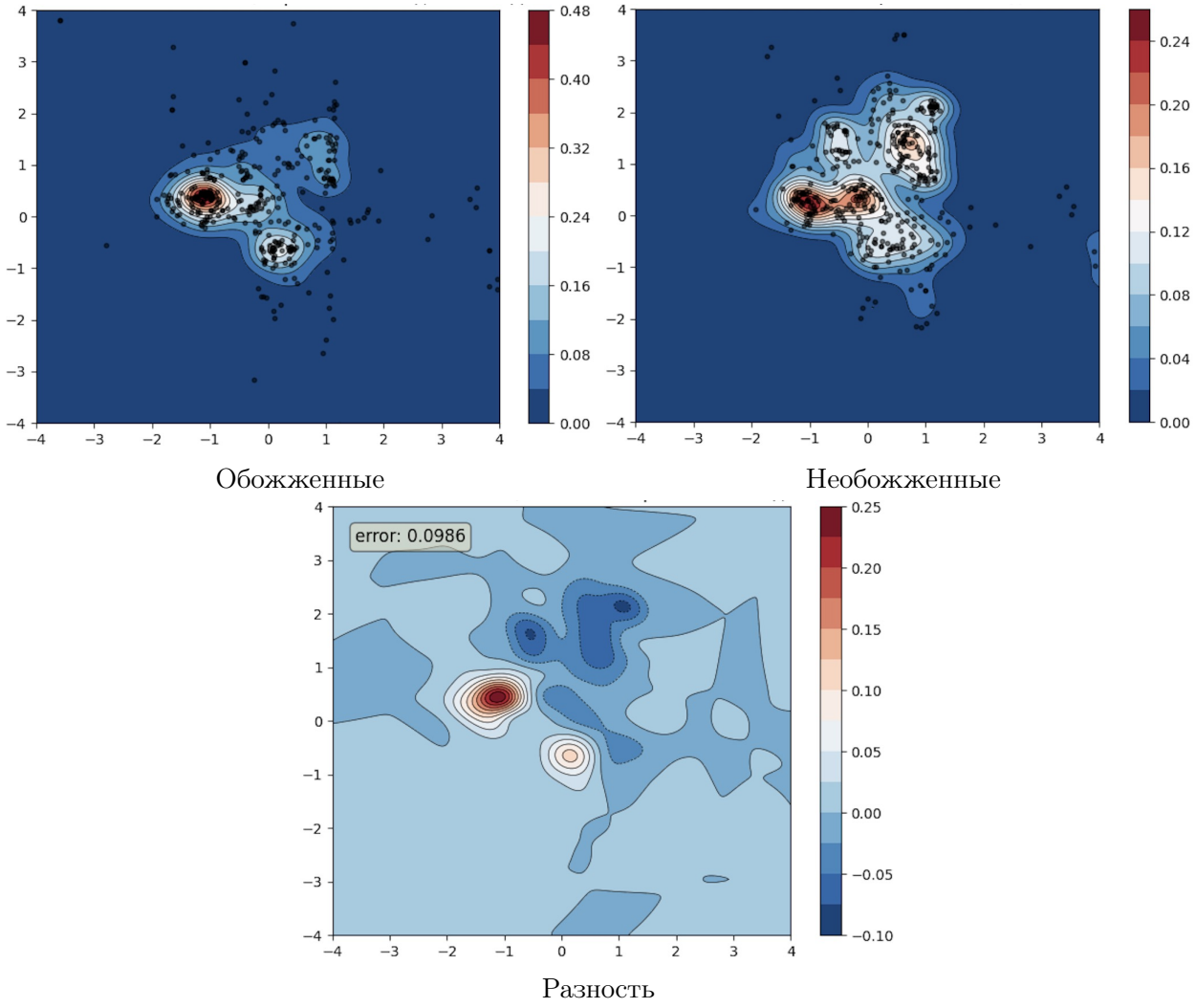


Рис. 1: Ядерные оценки плотностей находок в палево-желтом песке

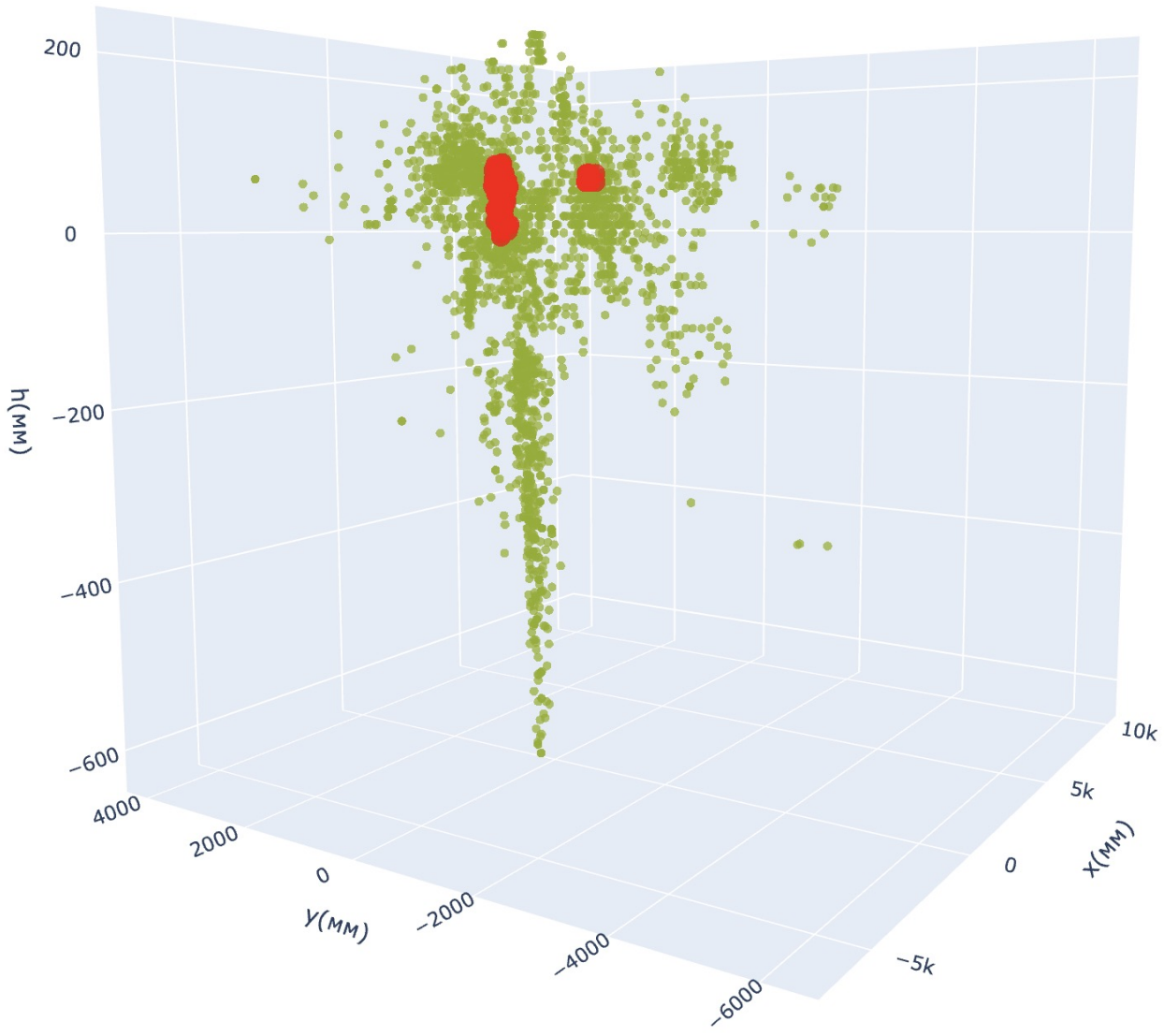


Рис. 2: Значимая концентрация обожженных находок в палево-желтом песке

В желтом песке значимым оказался участок термически необработанных находок, то есть темно-синяя область на нижнем графике рис. 3 свидетельствует о низкой концентрации термически обработанного материала по сравнению с остальными находками.
Пространственный анализ слоя желтого песка подтверждает результаты, полученные в двумерном случае. На рис. 4 отмечены находки, попавшие в пространственную область пониженной концентрации необожженных находок. В ней все находки, за исключением одной, являются термически необработанными.
Другие внешне выделяющиеся области концентрации (или разреженности), как следует из величин погрешностей, не могут быть признаны значимыми, и трактуются как случайности.

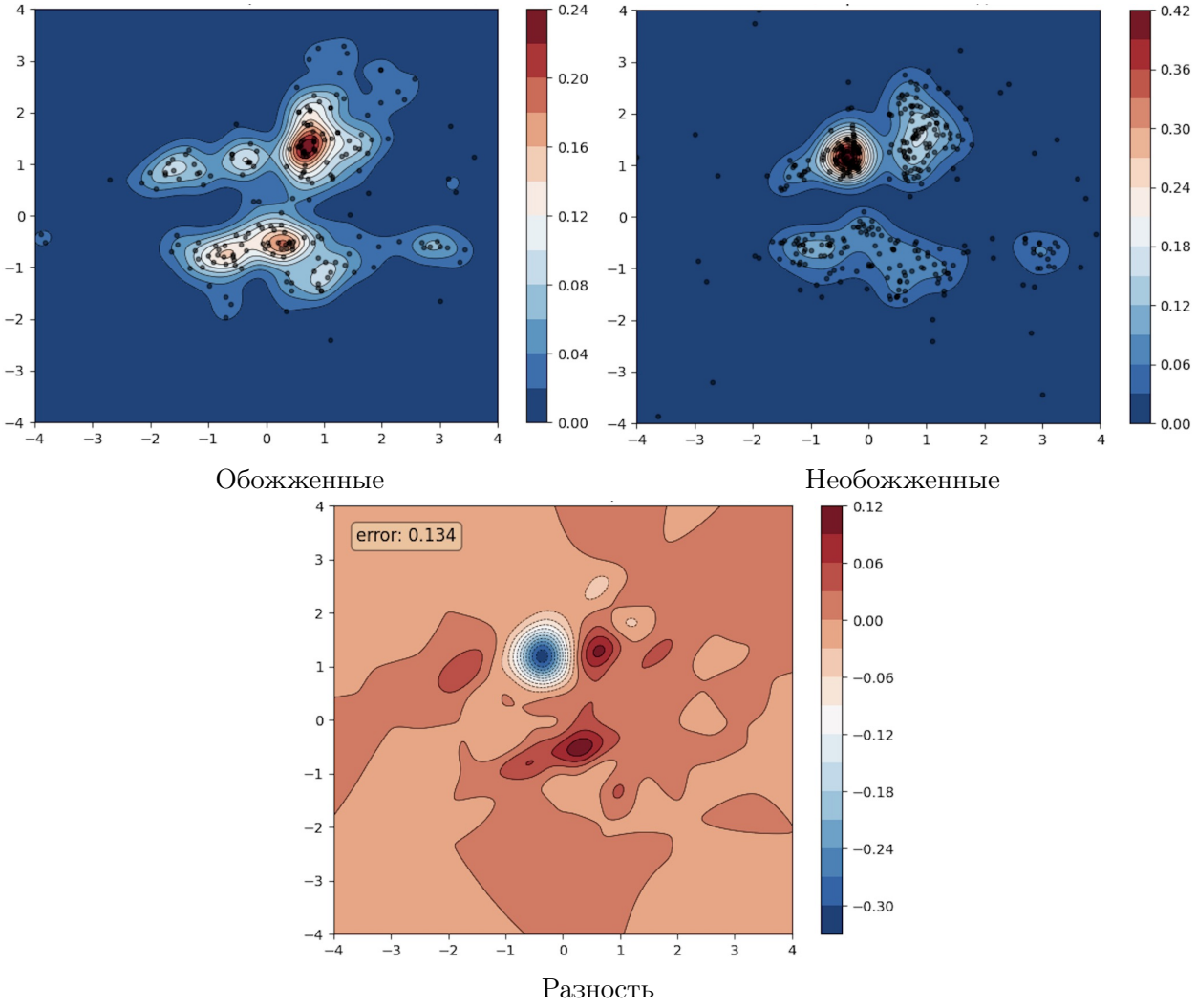


Рис. 3: Ядерные оценки плотностей находок в желтом песке

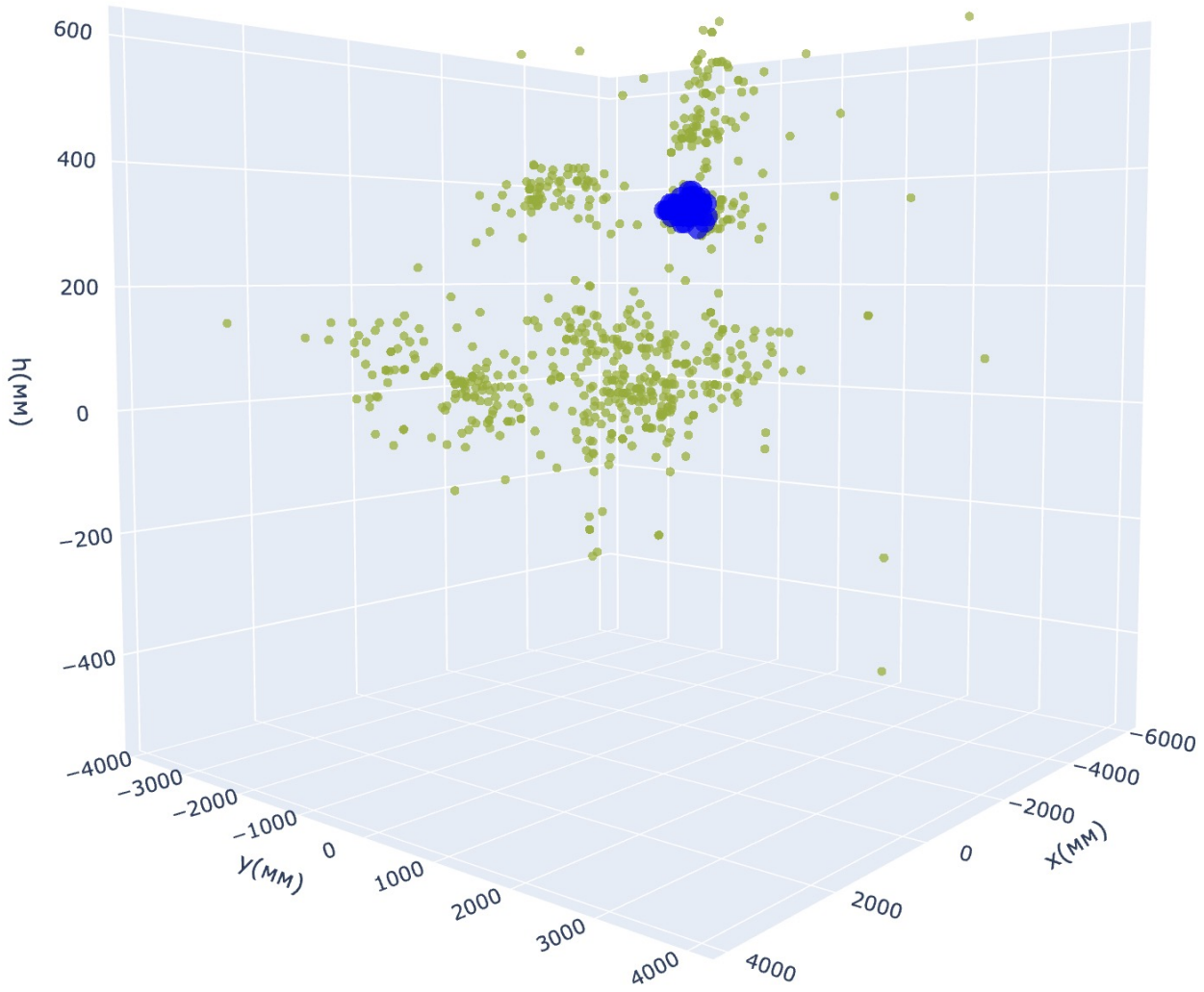
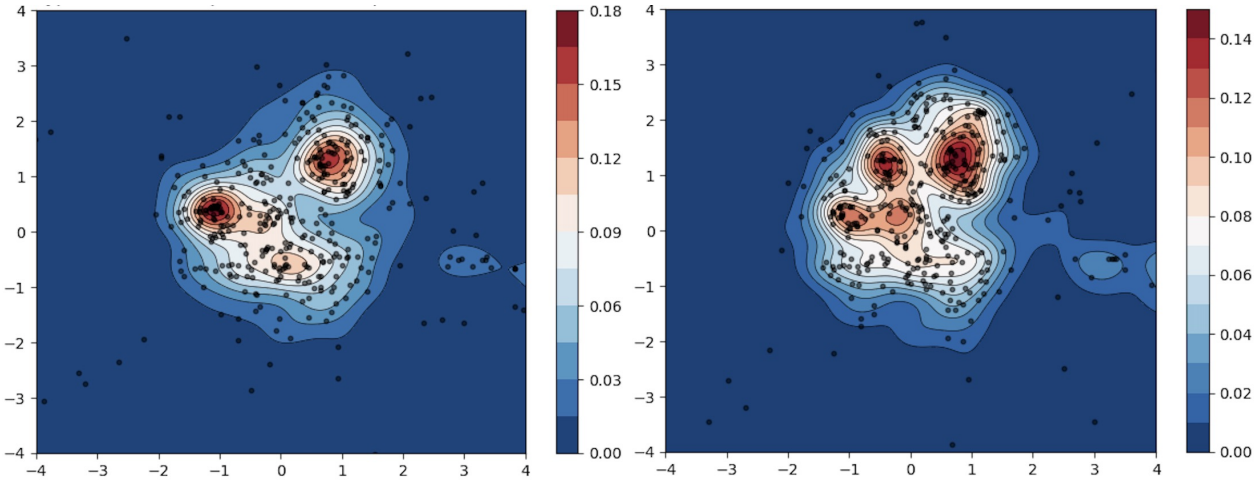


Рис. 4: Значимая концентрация необожженных находок в желтом песке

3.2.2 По категории

В этом разделе находки рассматривались в разрезе категорий.
При рассмотрении плоской проекции отходов расщепления (рис. 5) удалось выделить два участка: на одном преобладают обожженные находки, а на другом – необожженные. Эти участки по расположению совпадают с обнаруженными в предыдущем разделе.
Результаты трехмерного анализа приведены на рис. 6. Он подтверждает наличие участка значимой концентрации обожженного материала. Участок концентрации необожженного материала оказался крайне небольшим. Его яркая выраженность на рис. 5 связана с накоплением необожженного материала при проецировании данных на плоскость.



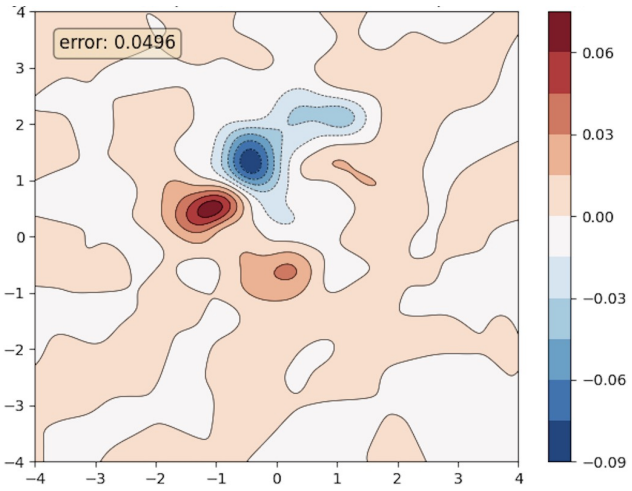


Рис. 5: Ядерные оценки плотностей отходов расщепления

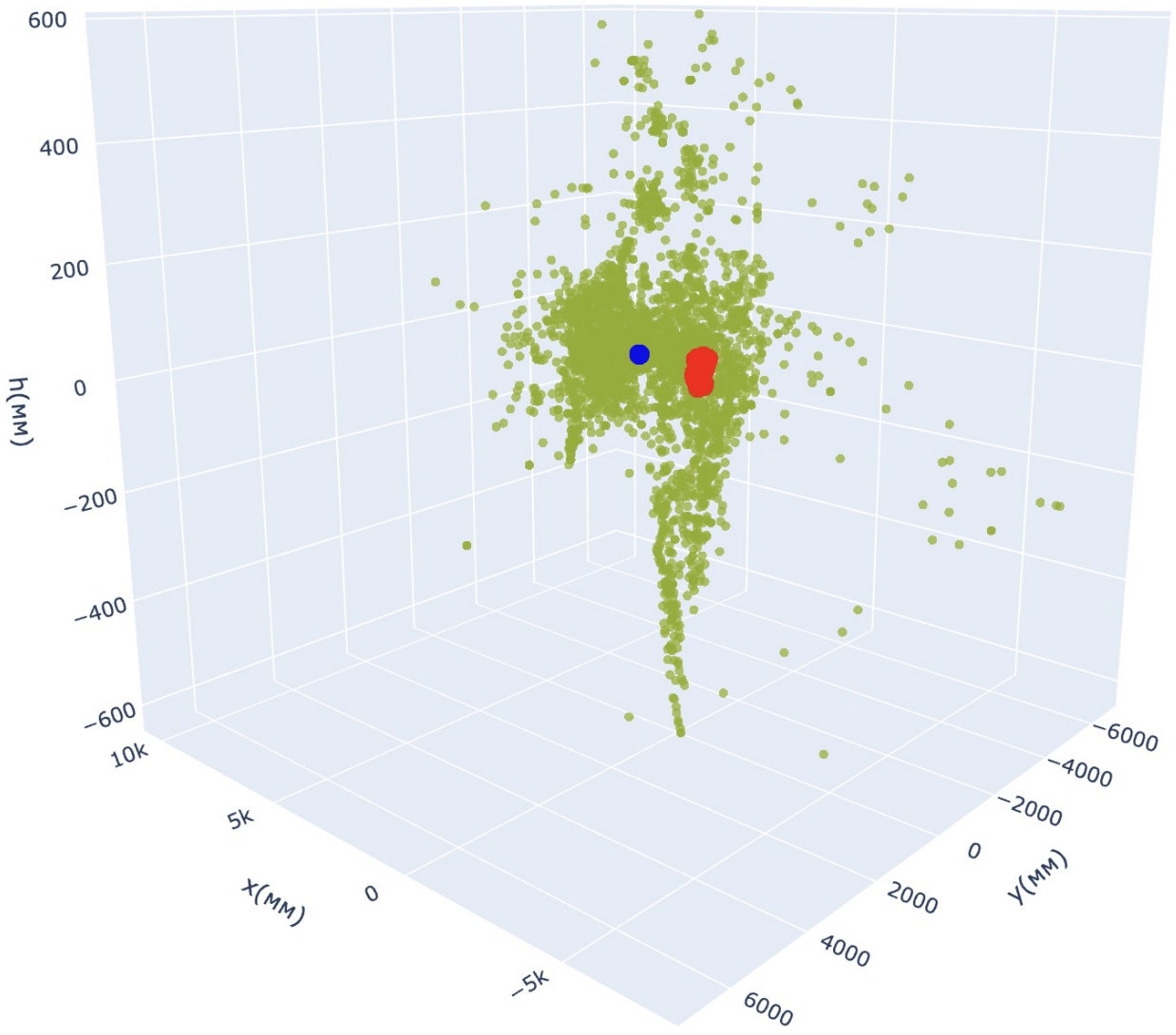
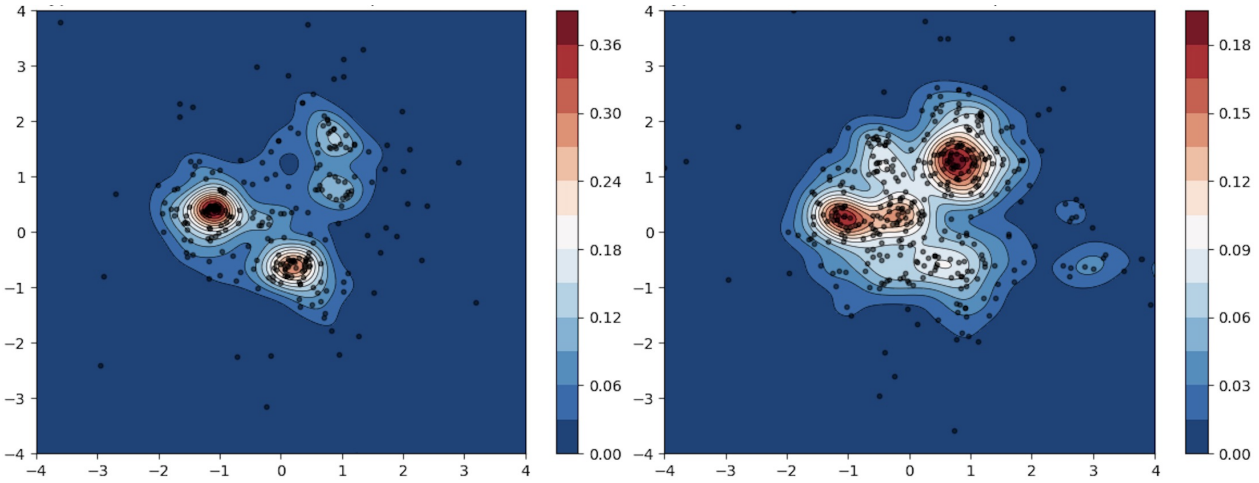


Рис. 6: Значимая концентрация отходов расщепления

Анализ плоской проекции пластинчатых сколов (рис. 7) дал две области значимого скопления термически обработанных находок. Трехмерный анализ (рис. 8) подтверждает их наличие. Более половины выделенных в пространстве находок оказались термически обработанными.



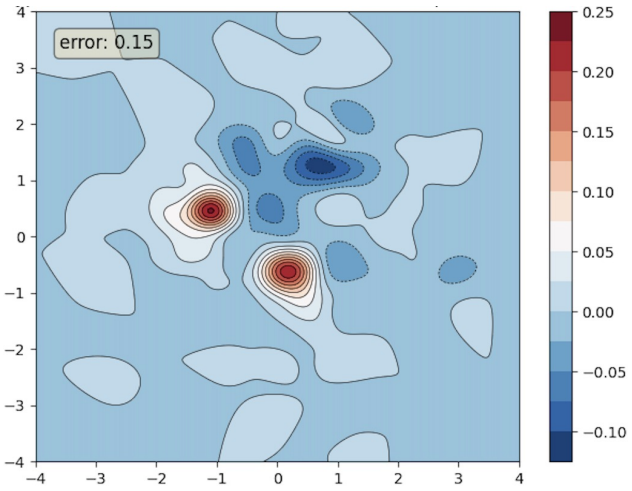


Рис. 7: Ядерные оценки плотностей пластинчатых сколов

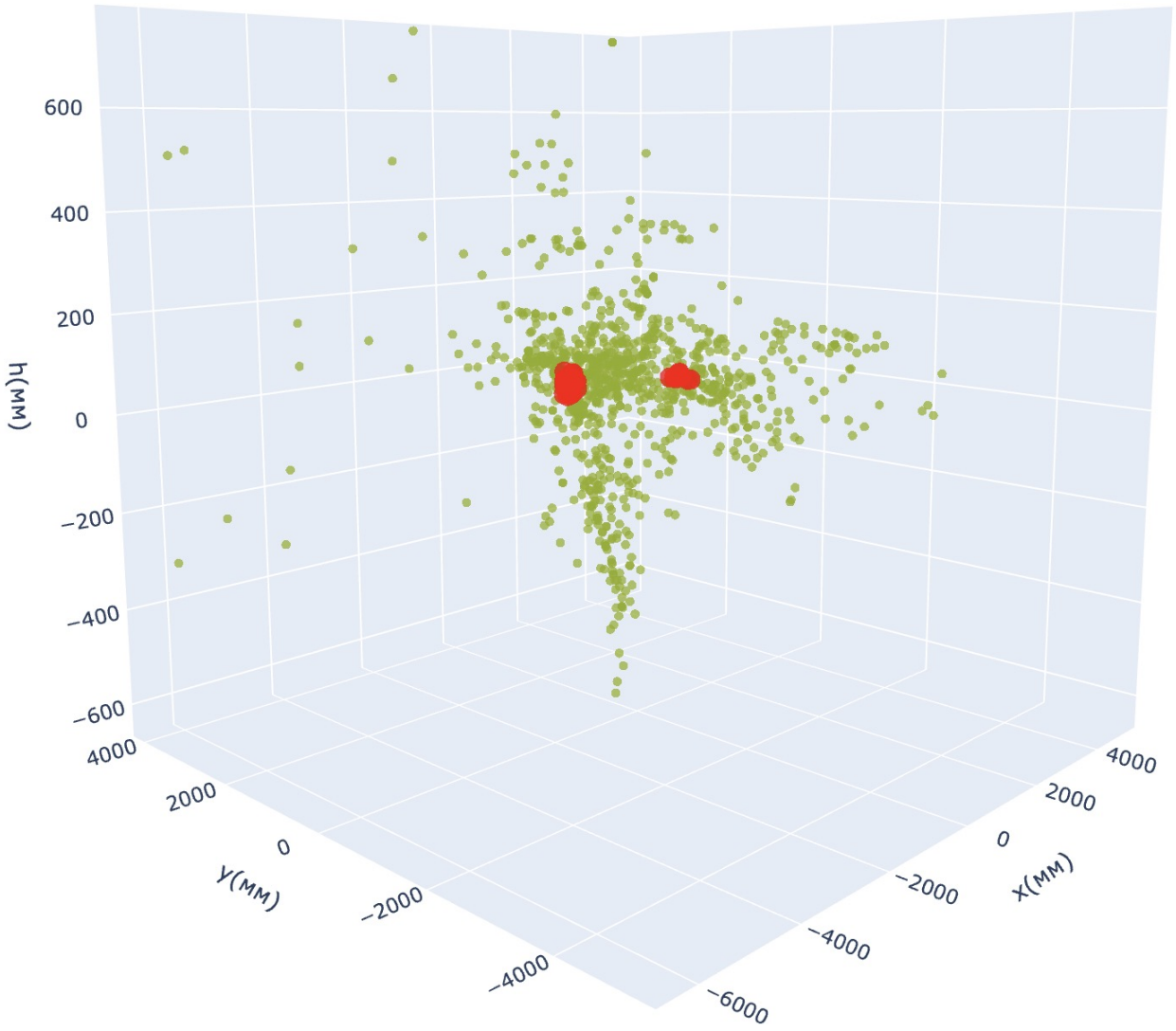
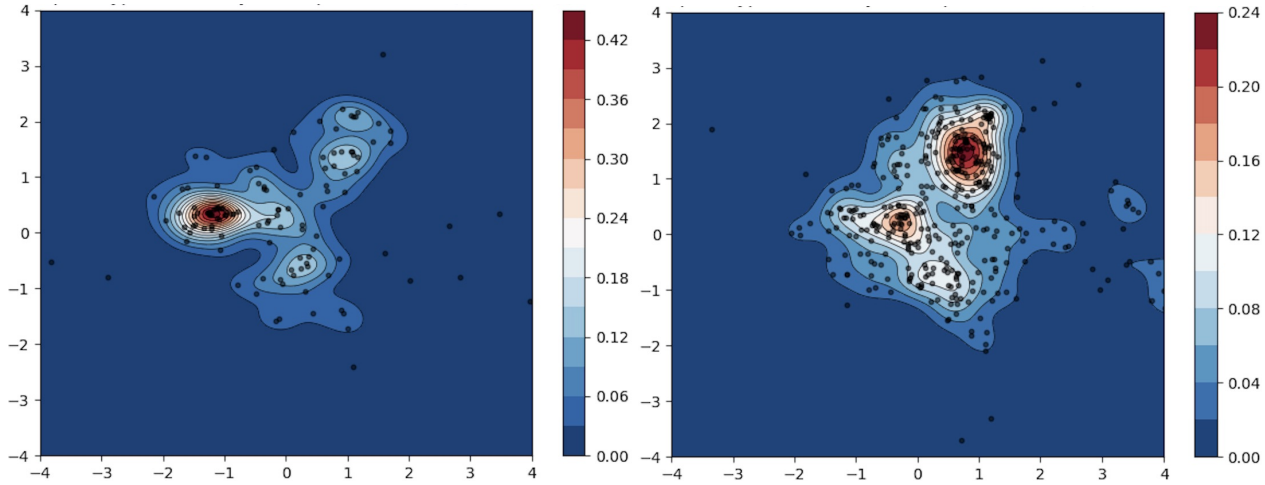


Рис. 8: Значимая концентрация обожженных пластинчатых сколов

Наконец, как видно из рисунков 9 и 10, термически обработанные чешуйки также концентрируются в ранее обнаруженном местоположении.



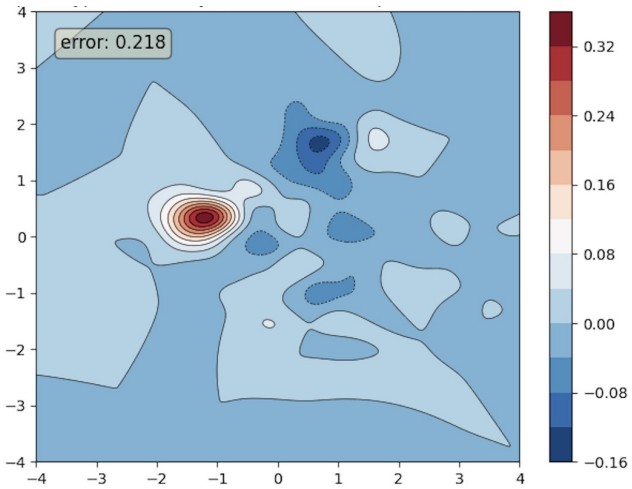


Рис. 9: Ядерные оценки плотностей чешуек

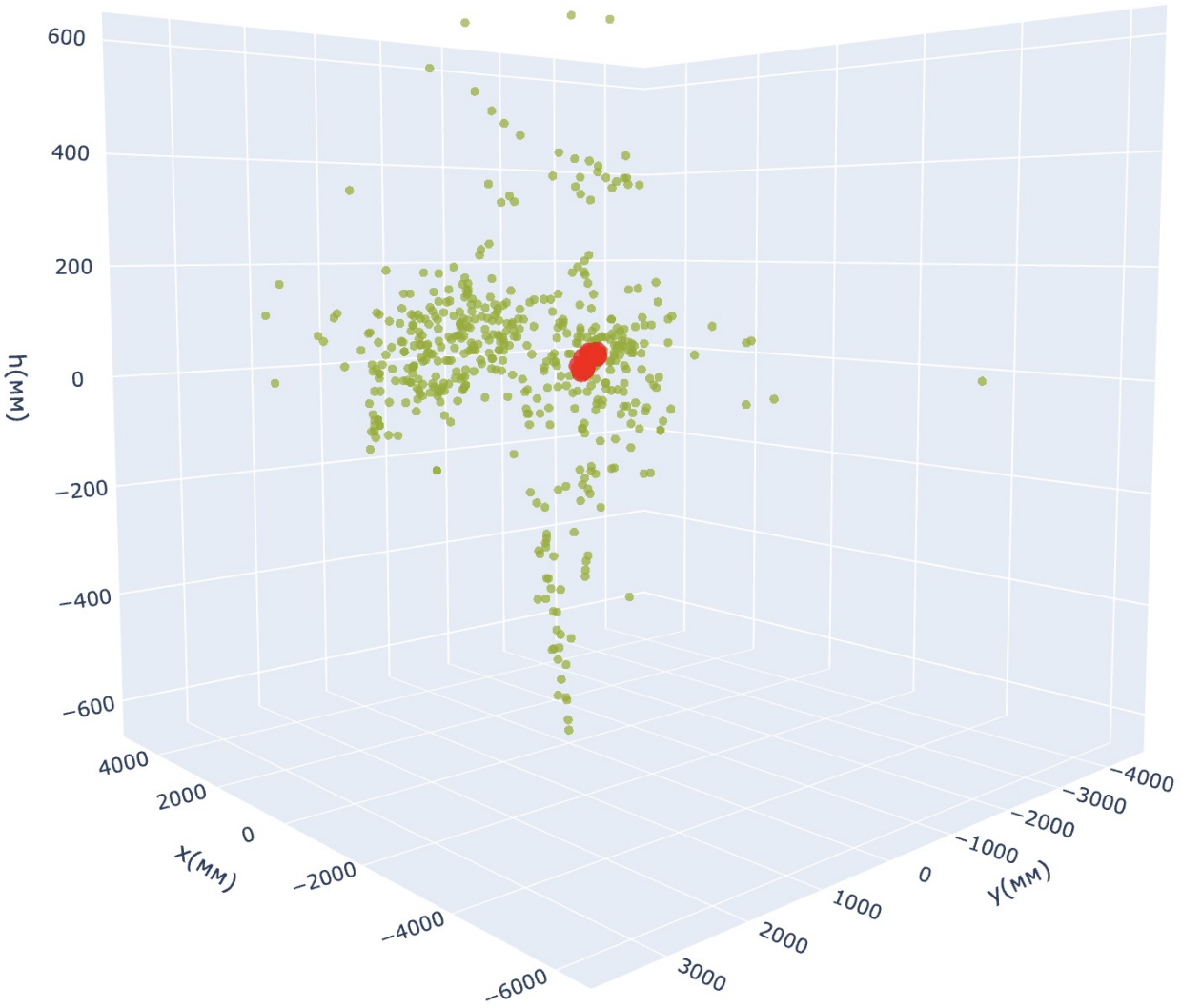


Рис. 10: Значимая концентрация обожженных чешуек

3.2.3 По типу сырья

При изучении находок в зависимости от типа сырья возникли трудности с малым количеством имеющихся данных. Анализ, подобный приведенным выше, удалось провести лишь в случае категории 5А (рис. 11). При этом он не выявил значимых отличий в расположении двух типов находок.

Сырье типов 2А и 10А встречалось лишь среди обожженных находок. Плотности их расположения приведены на рис. 12.

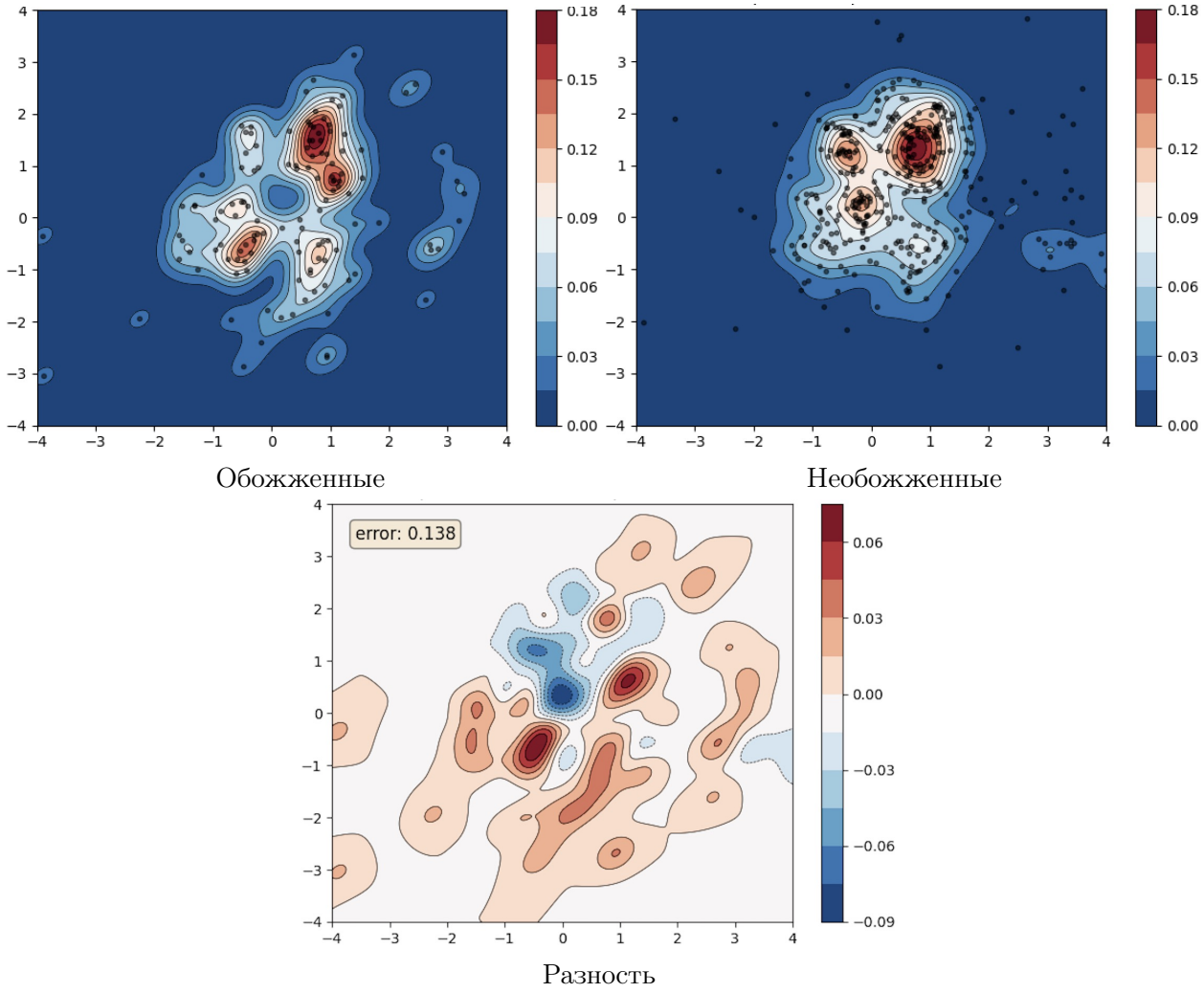


Рис. 11: Ядерные оценки плотностей находок из сырья 5А

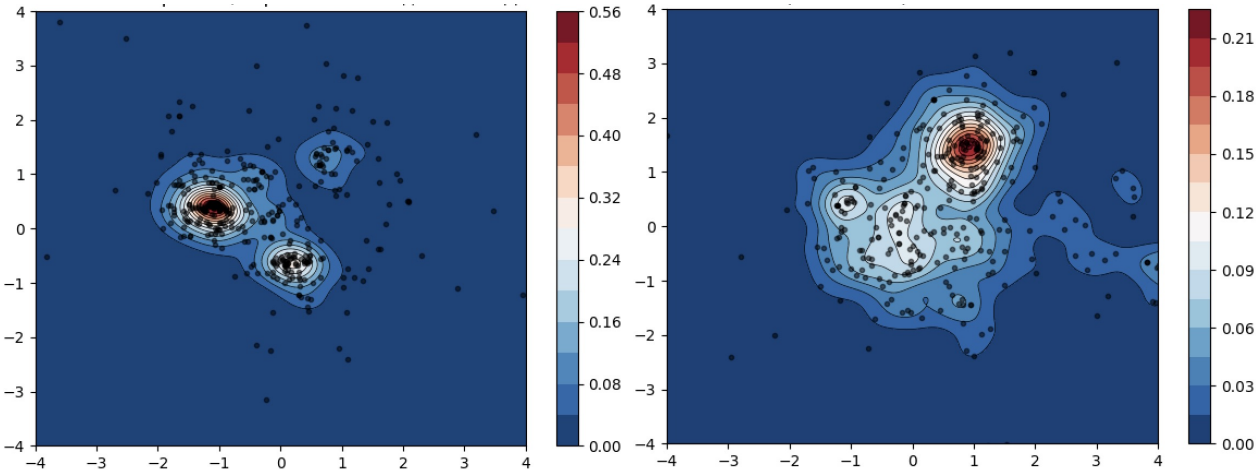


Рис. 12: Ядерные оценки плотностей обожженных находок

3.3 Замечание о ядерном методе оценивания

При применении ядерного метода оценивания плотности строились с использованием гауссовского ядра с шириной полосы пропускания равной 0.2. Существует множество ядер, которые могут быть использованы для построения оценок. Также можно варьировать параметр ширины полосы пропускания. Это сильно влияет на получаемые результаты: области концентрации получаются различными по форме и размерам, сливаются друг с другом и распадаются на более мелкие. Параметр пропускания равный 0.2 был выбран на основании нескольких экспериментов с учетом характера ожидаемого различия.

4 Сравнение методов

Первый метод ищет крупные области с фиксированным числом находок, где пропорция "обожженных/необожженных" существенно отличается.

Второй метод ищет небольшие области вокруг точки, где много точек, но обожженые и необожженные встречаются в особенной пропорции.

В частности, первый метод находит переферийные области большой площади с диспропорцией находок, а второй узкие скопления обожженных или необожженных находок.

5 Список литературы

[1] Yen-Chi Chen. A tutorial on kernel density estimation and recent advances, 2017.