АО «Научно исследовательский институт удобрений и инсектофунгицидов им. Я.В. Самойлова» (АО «НИУИФ»)

На правах рукописи

Юновидов Дмитрий Валерьевич

Программно-аппаратный рентгенофлуоресцентно-оптический комплекс для анализа сложных фосфорсодержащих удобрений

01.04.01 – Приборы и методы экспериментальной физики

Диссертация на соискание ученой степени кандидата технических наук

Научный руководитель
к.т.н., доц. Соколов Валерий Васильевич

Санкт-Петербург
2017

Содержание

1 Введение

Современное промышленное производство представляет сложный и многофакторный процесс, в котором участвуют три ключевых объекта: входящее сырье, промежуточные объекты и готовая продукция. Практика показывает, что без комплексного учета параметров и ключевых стадий производства невозможно добиться воспроизводимого и качественного конечного результата. Одним из наиболее информативных методов аналитического контроля является рентгенофлуоресцентный анализ (РФА). Данный метод широко распространен в аналитической и производственной практике, поскольку обладает целым рядом необходимых качеств: широкий диапазон определяемых концентраций (от 0,0001 до 100 мас. %); простая пробоподготовка; возможность анализа широкого спектра элементов – от бора до урана; экспрессность; многоэлементность; простота автоматизации; возможность использования в промышленных и полевых условиях; разнообразная приборная реализация (от портативных и дешевых энергодисперсионных переносных приборов, до сверхточных и чувствительных стационарных волновых систем).

В то же время, несмотря на развитость математического аппарата и разнообразие приборного парка, рассматриваемый метод не нашел широкого применения в промышленном производстве минеральных удобрений. Одна из возможных причин этого кроется в сложности промышленных объектов. Такие продукты обладают комплексной матрицей, что значительно увеличивает погрешность прямого анализа.

Однако развитие компьютерной техники, математического аппарата и аналитических методов позволяет накапливать и обрабатывать практически любые объемы информации для получения более подробного представления о протекающих процессах. С помощью алгоритмов предобработки данных, методов классификации и множественной регрессии становится возможным проведение многомерного анализа образцов. Приведенные методы анализа больших данных (АБД) отлично подходят для энергодисперсионного (ЭД) РФА, который обладает

высокой информативностью получаемого спектра. Дополнительной особенностью ЭД РФА является простота комбинирования получаемой информации с другими аналитическими и физическими методами контроля для комплексного описания производственного процесса.

Актуальность темы

Исследуемая тема актуальна в связи с практически полным отсутствием применения недорогого и эффективного метода ЭД РФА для контроля качества производимых минеральных удобрений. На сегодняшний день в одном из крупнейших производителей минеральных удобрений - холдинга «ФосАгро», на производстве работает только несколько приборов данного класса (в АО «Апатит» г. Кировск и в АО «Метахим», г. Волхов). Так же полностью отсутствуют базы данных анализируемых объектов. Практически нет отечественной методической и нормативной базы для данной группы методов и приборов на предприятиях.

Объект исследования

Сложные фосфорсодержащие минеральные удобрения и их параметры качества.

Цель работы

Создание программно-аппаратного комплекса на основе энергодисперсионного рентгенофлуоресцентного (РФ) спектрометра и оптического регистратора для повышения эффективности, экспрессности и надежности контроля качества выпускаемых сложных фосфорсодержащих удобрений (многофакторного мониторинга физических и химических свойств исследуемых проб).

Для достижения поставленной цели необходимо решить следующие **задачи**:

1. Выделить значимые химические и физические параметры для эффективного учета сложной матрицы комплексных фосфорсодержащих удобрений.
2. Разработать конкурентную и экономически эффективную аппаратную систему экспрессного получения физико-химической информации об объектах анализа.

3. Разработать и автоматизировать алгоритм выделения и расчета информативных признаков при контроле качества производимой продукции.
4. Создать прототип единой аналитической базы исследуемых объектов и обеспечить возможностью использования методов анализа больших данных.
5. Теоретически обосновать и разработать схему комплексного анализа сложных фосфорсодержащих удобрений для минимизации потерь сырья и энергоресурсов при переходе с одной производимой марки на другую.
6. Обеспечить удобство пользования данным методом в заводских лабораториях (упрощение программной, процедурной и аналитической части комплекса).

Научная новизна.

В результате выполнения диссертационной работы:

1. На основании полученных экспериментальных данных разработана ранее не применявшаяся программно-аппаратная схема оптического анализатора с ЭД РФ спектрометром для многофакторного экспрессного анализа свойств сложных фосфорсодержащих удобрений.
2. Разработаны и автоматизированы экспрессные методы определения содержания различных химических элементов, типа, марки, фракционного состава и степени обработки кондиционирующими добавками (к.д.) анализируемых объектов.
3. На основании данных рентгенофлуоресцентно-оптического комплекса создана оригинальная аналитическая база данных физических и химических свойств исследуемых объектов, позволяющая увеличить точность и быстродействие измерений.
4. Показана возможность проведения регрессионного и классификационного анализа марок выпускаемых удобрений по всем основным питательным элементам (N, P, K) и серы в широком концентрационном диапазоне.
5. На основе ЭД РФ спектрометра и составленной базы данных предложен способ определения азота в минеральных удобрениях, прямое детектирование которого методом ЭД РФА невозможно.

6. Создано алгоритмическое и программное обеспечение для разработанного аппаратного комплекса, обеспечивающее автоматический расчет аналитических сигналов, поиск корреляций и статистический анализ больших массивов данных.
7. Разработан способ определения фракционного состава запрессованных проб для ЭД РФА с использованием системы оптического контроля.

Практическая значимость.

Разработанные методы контроля и приборы на их основе используются при производстве сложных фосфорсодержащих удобрений на предприятиях холдинга «ФосАгро». Разработанные алгоритмы обработки данных использованы в отечественных ЭД РФ спектрометрах производства АО «Научные приборы».

1. Для реализации метода измерения физико-химических свойств готовой продукции разработан программно-аппаратный комплекс, который позволил увеличить чувствительность, точность и быстродействие исследования качества выпускаемых сложных фосфорсодержащих удобрений.
2. Создан оригинальный прототип программно-аппаратного комплекса для оценки качества производимой продукции и экспресс-анализа на химический состав по основным питательным элементам (N, P, K), сере и фракционному составу. Предложенное оборудование имеет широкие перспективы для анализа промышленных объектов как в лаборатории, так и непосредственно в производственных условиях.
3. Разработана и реализована оригинальная методика контроля таких физических свойств гранулированных минеральных удобрений, так гранулометрический состав и качество обработки кондиционирующими добавками.
4. Предложена схема комплексного анализа сложных фосфорсодержащих удобрений для минимизации потерь сырья и энергоресурсов при переходе с одной производимой марки на другую.

Достоверность научных положений и выводов подтверждается соответствием разработанных физико-математических моделей и теоретических расчетов с результатами большого объема экспериментальных исследований.

Положения, выносимые на защиту.

- Программно-аппаратный комплекс на основе ЭД РФ-спектрометра и оптического регистратора, позволяющий проводить многофакторный экспрессный анализ сложных фосфорсодержащих удобрений.
- Алгоритм создания базы данных физико-химических свойств промышленных объектов, таких как: содержание различных химических элементов, тип, марка, фракционный состав и степень обработки кондиционирующими добавками.
- Оригинальная математическая модель экспрессного комплексного анализа пробы для определения:
  - физических свойств: тип, максимальная фракция и наличие кондиционирующей добавки.
  - химического состава и марки выпускаемых удобрений по всем основным питательным элементам (N, P, K) и сере, включая азот, прямое определение которого методом ЭД РФ анализа невозможно.
- Алгоритмическое и программное обеспечение «DSpectra» для разработанного аппаратного комплекса, обеспечивающее автоматизированный расчет аналитических сигналов, поиск корреляций и статистический анализ больших массивов данных.

Апробация работы и публикации.

Результаты диссертационной работы использованы при решении аналитических задач производства минеральных удобрений в лабораториях промышленных объектов холдинга «ФосАгро», АО «НИУИФ» и для совершенствования аналитических приборов АО «Научные приборы».

Основные положения диссертационной работы доложены на конференциях и семинарах:

- «2 съезд аналитиков России», г. Москва, 2012 г.;
- научно-практический семинар «Роль аналитических служб в обеспечении качества минеральных удобрений и серной кислоты», г. Москва, 2013 и 2014 г.г.;
- «VIII международная конференция по рентгеноспектральному анализу», г. Иркутск, 2014 г.;
- научная конференция молодых ученых «Ломоносов», г. Москва, 2014 и 2015 г.г.,
- семинар «Новое в теории и практике рентгенофлуоресцентного анализа. Развитие программного и методического обеспечения рентгеновских аналитических приборов производства АО «Научные приборы»», г. С. Петербург, 2016 г.
- «3 съезд аналитиков России», г. Москва, 2017 г.

В соответствии с паспортом специальности (01.04.01 – «Приборы и методы экспериментальной физики») в диссертационной работе проведена разработка методов измерений физических величин, позволяющих существенно увеличить точность, чувствительность и быстродействие систем контроля качества выпускаемой продукции. Реализована автоматизация физического эксперимента. Разработан малогабаритный и эффективный прибор для получения комплексной физической и химической информации о качестве минеральных удобрений.

Публикации. По материалам диссертации опубликовано 8 печатных работ, из них 2 в журналах, входящих в Перечень ведущих рецензируемых научных журналов и изданий ВАК РФ, 9 тезисов докладов на всероссийских и международных конференциях и семинарах, 2 патента РФ на стадии оформления, одно свидетельство о государственной регистрации программы для ЭВМ.

1. Юновидов Д.В., Соколов В.В., Бахвалов А.С. Метод оценки влияния стадий пробоподготовки NPKS удобрений на результаты рентгенофлуоресцентного анализа по спектру пробы // Заводская лаборатория. Диагностика материалов. 2017. Т. 83, № 9 с. 15-21.
2. Юновидов Д.В., Соколов В.В., Бахвалова Е.В, Донских В.А. Разработка стандартного образца апатитового концентрата. Эффективный контроль

однородности с помощью рентгенофлуоресцентных методов анализа // ГИАБ. 2016. № 7. с. 131-144.

3. Свидетельство № 2017617704 Российская Федерация. Программа «DSpectra» / Юновидов Д.В.; заявитель и правообладатель Юновидов Д.В. - № 2017614722; заявл. 19.05.2017; зарегистрировано в Реестре программ для ЭВМ 11.07.2017 – [1] с.
4. Юновидов Д.В., Соколов В.В., Осколок К.В., Болотоков А.А.. Рентгенофлуоресцентное определение церия в экстракционной фосфорной кислоте и фосфатных концентратах // Мир серы, N, P и K. 2012. №4. с. 10-13.
5. Юновидов Д.В., Соколов В.В., Осколок К.В., Болотоков А.А.. Рентгенофлуоресцентное определение редкоземельных элементов после сорбционного выделения и концентрирования из экстракционной фосфорной кислоты // Фосфатное сырье: производство и переработка. 2013. с. 147 – 151.
6. Юновидов Д. В., Эль-Салим С.З., Осколок К.В. Восстановление спектра гомогенной системы по временным зависимостям интенсивностей линий в зарождающейся и развивающейся гетерогенной системе на примере экстракционной фосфорной кислоты // VIII Всероссийская конференция по рентгеноспектральному анализу. Иркутск, 22 - 26 сентября 2014 г. Тезисы докладов. — Иркутск. Институт земной коры СО РАН, 2014. - С. 139–139.
7. Юновидов Д. В., Эль-Салим С. З., Осколок К. В. Техника виртуального эксперимента и её применение для количественного рентгенофлуоресцентного анализа экстракционной фосфорной кислоты // VIII Всероссийская конференция по рентгеноспектральному анализу. Иркутск, 22 - 26 сентября 2014 г. Тезисы докладов. — Иркутск. Институт земной коры СО РАН, 2014. - С. 140–140.
8. Юновидов Д.В., Ребрикова А.Т., Осколок К.В., Соколов В.В. Рентгенофлуоресцентное определение технологически важных элементов в экстракционной фосфорной кислоте // Второй съезд аналитиков России. Москва, 23 - 27 сентября 2013 г. Тезисы докладов. - Москва. - С. 289–289.

9. Юновидов Д.В., Ребрикова А.Т., Осколок К.В., Соколов В.В. Техника виртуального эксперимента для количественного рентгенофлуоресцентного анализа экстракционной фосфорной кислоты // Второй съезд аналитиков России. Москва, 23 - 27 сентября 2013 г. Тезисы докладов. - Москва. - С. 290–290.

Структура и объем работы. Содержание диссертации изложено на 187 страницах и состоит из введения, шести глав, заключения, приложения и списка литературы, содержащего 116 наименований. Работа содержит 50 таблиц и иллюстрирована 65 рисунками.

2 Обзор литературы

На сегодняшний день понятия сельское хозяйство и минеральные удобрения неразрывно связаны друг с другом. Добыча фосфатной руды переживает интенсивный подъем, с ожидаемым пиком через 20 – 30 лет [1,2]. Различные виды удобрений, получаемые из обогащенной фосфатной руды, используются повсеместно для выращивания разнообразных посевных культур и увеличения объемов продукции сельского хозяйства. Наиболее значимыми элементами для жизни растения после углерода, кислорода и водорода являются азот, фосфор и калий [2,3] – именно они названы «основными питательными веществами» и составляют основу минеральных удобрений, используемых в сельском хозяйстве. Наиболее крупным производителем минеральных удобрений на территории России является холдинг компаний «ФосАгро», обладающей собственной сырьевой базой и полным циклом производства современных агрохимикатов. Продукция данного производителя является объектом анализа на протяжении всего настоящего исследования.

2.1 Фосфатное сырье

Фосфатные руды являются важнейшим источником фосфора и различных примесных элементов в минеральных удобрениях. В промышленных целях в основном, используют кальциевые фосфаты: апатиты и фосфориты [1,3,4]. Фосфориты бывают разных типов (таблица 2.1) [1]:

Таблица 2.1. Типы фосфоритов

Тип отложений	**Примерный процент от общих залежей, %**
морские	75
магматические	15 - 20
метаморфические	
в результате выветривания	
биогенные	2 - 3

Наиболее распространенной формой фосфора являются соединения группы фосфатов кальция под общим названием «апатиты», описываемые простейшей формулой $3Ca_3(PO_4)_2 \times CaX_2$, где X как правило фтор или, гидроксильная группа, реже хлор [1,4–7]. Часть атомов кальция в кристаллической решетке апатита достаточно часто изоморфно замещаются на более тяжелые атомы, например, такие как: уран, стронций и редкоземельные элементы (РЗЭ) [6–13]. Помимо фторапатита в состав апатитовых руд входят примеси нефосфатных минералов, таких как нефелин $(Na, K)_2O \times Al_2O_3 \times 2SiO_2 \times 2H_2O$), эргин $[Na_2Fe](SiO_3)_2$, титано-магнетит $Fe_3O_4 \times FeTiO_3 \times TiO_2$, ильменит $FeTiO_3$, сфен $CaTiSiO_5$ и др., чем объясняется широкий диапазон примесных элементов, обусловленный природой апатита и местом его добычи.

Существует ограниченный набор стран, добывающих фосфориты и имеющих достаточные запасы данного вида минерала (таблица 2.2) [3,5,14].

Таблица 2.2. Основные страны добывающие фосфориты. Для удобства метрические тонны пересчитаны на тонны.

Страна	**Уровень добычи, 10^4 т. в год**	**Ориентировочные запасы, 10^7 т.**
Китай	8,9	1
США	2,92	0,4
Марокко	2,8	2,1
Россия	1,13	0,1

При этом каждое из месторождений обладает своим уникальным химическим составом, зависящим от типа фосфоритов (осадочные или магматические) и географического положения. В таблице 2.3 приведен примерный состав концентрированной фосфатной руды основных месторождений [6,15,16].

Таблица 2.3. Химический состав фосфатной руды

Страна	Область	Химический состав	
		форма пересчета	концентрация, мас. %
Китай	Вухан	P_2O_5	23,98
		CaO	32,02
		F	3,60
		SiO_2	22,14
		Fe_2O_3	2,29
		Al_2O_3	4,11
		MgO	2,11
США	Северная Каролина	P_2O_5	29,7
		CaO	47,4
		Cl	0,015
		F	3,53
		SiO_2	1,73
		Fe_2O_3	0,79
		Al_2O_3	0,53
		MgO	0,79
		Na_2O	0,98
		K_2O	0,17
		CO_2	4,18
		Орг. С	1,38
		Общая S	1,1
США	Флорида	P_2O_5	31,2
		CaO	45,0
		Cl	0,05
		F	3,60
		SiO_2	9,48
		Fe_2O_3	1,33
		Al_2O_3	1,76
		MgO	-
		Na_2O	0,89
		K_2O	0,11
		CO_2	3,48
		Орг. С	2,18
		Общая S	1,05
Марокко	Сахара	P_2O_5	34,2
		CaO	50,3
		Cl	0,02
		F	3,8
		SiO_2	-
		Fe_2O_3	0,22
		Al_2O_3	0,48
		MgO	0,12
		Na_2O	-
		K_2O	-
		CO_2	2,7
		Орг. С	0,06
		Общая S	-

Марокко	Сафи	P_2O_5 CaO Cl F SiO_2 Fe_2O_3 Al_2O_3 MgO Na_2O K_2O CO_2 Орг. С Общая S	32,4 49,9 0,02 4,1 2,85 0,70 0,40 0,70 0,90 0,10 4,1 - 0,20
Россия	Хибины	P_2O_5 CaO F SiO_2 Fe_2O_3 Al_2O_3 MgO Na_2O K_2O CO_2	39,4 51,5-52,0 3,0-3,1 1,0-1,5 3,0 3,0 0,2-0,4 0,8-1,0 0,8-1,0 -

Как правило, промышленно значимой признается руда с относительно высоким содержанием фосфора – на уровне 30 % и более, соотношением CaO / P_2O_5 не более 1,6 и содержанием MgO менее 1% [5], для чего используют предварительное очищение ее от примесей – флотацию. Именно эти элементы, наряду с основными питательными являются маркерами технологического процесса.

Большое количество статей посвящено улучшению способов флотации и обогащения бедных руд [4,5,12,17–22]. Учитывая все возрастающее потребление фосфатного сырья и его невозобновляемость, в скором времени технологические требования могут измениться, а удельное количество примесей в исходном сырье и конечном продукте возрасти. На наш взгляд, данные предпосылки так же должны учитываться при внедрении новых методов аналитического контроля на предприятиях.

Для отечественного производства минеральных удобрений, как правило, используются апатиты (таблица 2.3). После обработки и концентрации руды

получают апатитовый концентрат с содержанием P_2O_5 не менее 38 %; из примесных элементов как правило присутствуют: Fe, Mg, Sr, Al, F, Si и.т.д. Предприятие АО «Апатит» полностью обеспечивает «ФосАгро» апатитовым концентратом для производства фосфорсодержащих удобрений. При этом рудная база содержит значительные запасы оксида алюминия (Al_2O_3) и более 41% всех российских запасов редкоземельных элементов. Именно хибинский апатит является основой отечественной промышленности минеральных удобрений и далее в работе термины «фосфориты» и «хибинский апатитовый концентрат» используются как синонимы.

В свою очередь, под термином «минеральные удобрения» понимают неорганические соли, включающие в себя набор питательных элементов: азот, фосфор, серу, калий; и микроэлементы: магний, бор, цинк, кальций и железо [3,13]. Основными питательными являются азот, фосфор и калий, которые присутствуют практически в любой марке промышленно производимых комплексных минеральных удобрений. Однако не менее значимыми являются и микроэлементы. Функции каждого из макро- и микро- элементов специфичны и не заменимы [7].

В дальнейшем будут рассматриваться именно комплексные фосфорсодержащие минеральные удобрения.

2.2 Процесс производства минеральных удобрений. Объекты аналитического контроля

Как уже отмечалось выше – объектом аналитического контроля выбраны сложные фосфорсодержащие минеральные удобрения. Поскольку в АО «Апатит» холдинга «ФосАгро» добыча сырья для производства удобрений происходит из одного месторождения, объект является достаточно чистым и стабильным. Требования к контролю и концентрации основных элементов апатитового концентрата приведен в таблице 2.4 [23].

Таблица 2.4. Требования аналитического контроля апатитового концентрата.

Элемент	Концентрационный диапазон, %	Требуемая точность, %
CaO	40,0 – 53,0	0,04 – 0,4
P_2O_5	38,0 – 40,0	0,2
MgO	2,2 – 4,8	0,2
As	0,00006 – 0,008	0,00002 – 0,0008
Cl	0,003 – 0,016	–
H_2O	1,0	0,5

Далее из апатитового концентрата получают экстракционную фосфорную кислоту (ЭФК). На сегодняшний день для получения ЭФК и различных минеральных удобрений в большинстве случаев используют кислотное разложение апатита серной кислотой [3,4,16,20]. Данный метод заложен в основу работы холдинга «ФосАгро» и других предприятий отечественной промышленности по производству минеральных удобрений, именно его мы в дальнейшем и будем рассматривать в представленной работе.

В результате кислотной переработки получают фосфорную кислоту, которую используют в качестве сырья для получения различных марок фосфор-содержащих удобрений.

Общий химизм производства минеральных удобрений в России и мире [2–4,16] представлен на рисунке 2.1.

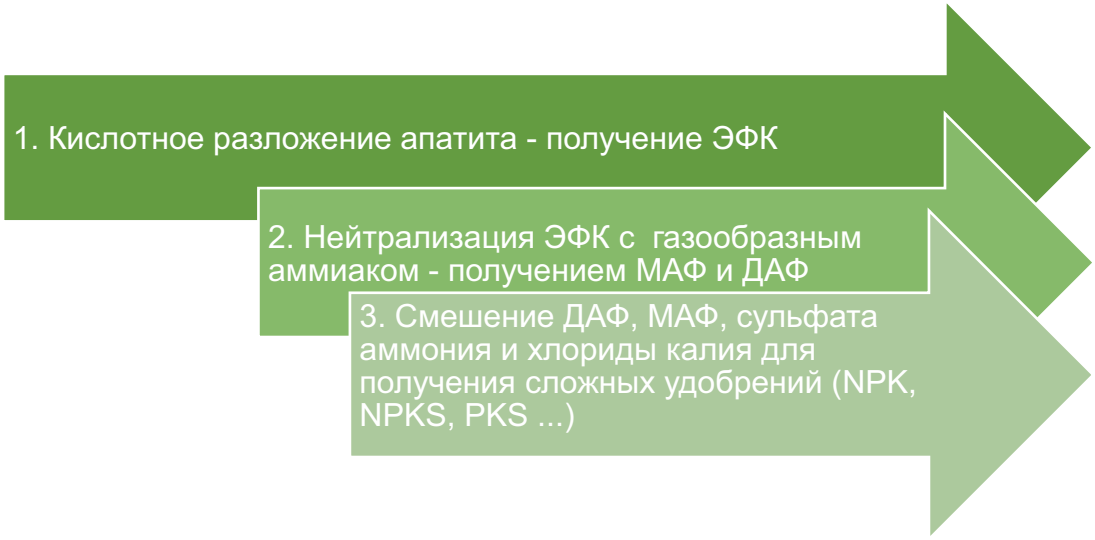


Рисунок 2.1. Химизм процесса производства минеральных удобрений

При этом стадии 1 - 3 могут быть представлены следующими химическими реакциями:

1. $Ca_5F(PO_4)_3 + 5H_2SO_4 + nH_3PO_4$ (оборотная) $+ mH_2O \rightarrow (n+3)H_3PO_4 + 5CaSO_4 \cdot mH_2O\downarrow + HF$
2. $H_3PO_4 + NH_3 \rightarrow NH_3H_2PO_4$ или $H_3PO_4 + NH_3 \rightarrow (NH_3)_2HPO_4$

На первой стадии процесса переработки апатита получают экстракционную фосфорную кислоту (ЭФК) которая так же является источником различных примесных элементов в минеральных удобрениях [24]. Фосфорная кислота является полупродуктом в производстве минеральных удобрений и может производиться кислотной экстракцией из апатита по дигидратнной, полугидратной и смешанной схемам, которые отличаются получаемой формой гипса, матричным составом и физическими параметрами процесса (таблица 2.5) [2–4,6,16].

Таблица 2.5. Сравнение процессов экстракции апатита.

Тип процесса	**Концентрация получаемой фосфорной кислоты, %**	**Степень извлечения P_2O_5 из апатита, %**	**Время процесса, ч**	**Температура процесса, 0С**
Дигидратный	28 - 30	98	6 - 8	75 – 80
Полугидратный	34 - 35	96,5	4 – 6	94 - 100

Обе схемы приводят к получению неупаренной ЭФК относительно низкой концентрации, которая в последствии подлежит упарке до 52 - 54 %. Оба процесса обладают примерно одинаковыми технологическими схемами, хотя полугидратный процесс является более сложным по поддержанию режима и обладает иной матрицей ЭФК. Типовая схема процесса экстракции фосфоритов приведена на рисунке 2.2.

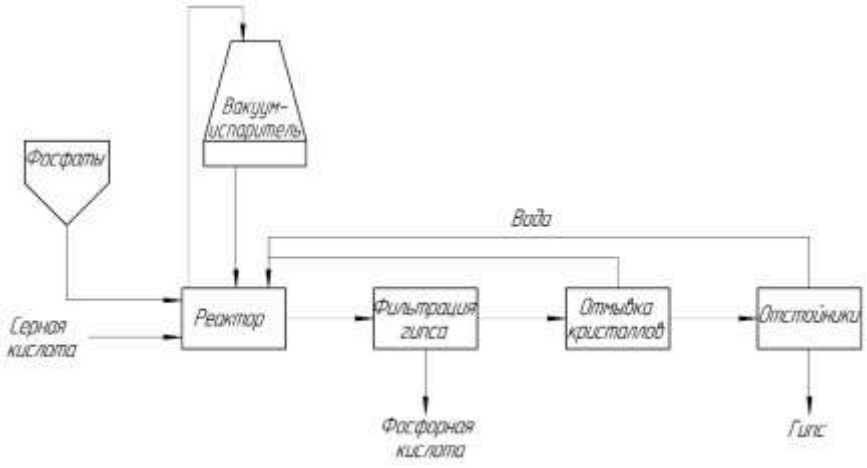


Рисунок 2.2. Схема экстракции ЭФК из фосфатного сырья.

В реальных условиях получаемая фосфорная кислота загрязнена примесями различных металлов (в том числе и тяжелых) а так же серной, кремнефтористой и зачастую фтористо-водородной кислотами [4,16]. Основные из побочных реакций, которые приводят к повышенному потреблению серной кислоты и уменьшению выхода целевой реакции:

- $Fe_2O_3 + H_2SO_4 \rightarrow Fe_2(SO_4)_3 + H_2O$
- $Al_2O_3 + H_2SO_4 \rightarrow Al_2(SO_4)_3 + H_2O$
- $MgO + H_2SO_4 \rightarrow MgSO_4 + H_2O$

Это определяет технологические требования к перерабатываемым фосфатам (предельное допустимое содержание тех или иных примесей) [1,5,13], жесткие требования к протеканию процесса (контроль сернокислотного режима, плотности и т.д.) [2,3] и коррозионную активности получаемой ЭФК [4,16]. Так, согласно рассмотренным источникам, основными элементами – маркерами качества, подлежащими аналитическому контролю в ЭФК являются:

- S – сернокислотный режим: влияет на скорость реакции разложения апатита и формирования крупных кристаллов фосфогипса для качественной отмывки;
- Si, Al – элементы понижающие коррозионную активность за счет реакции с HF;
- F – коррозионные агенты;
- Fe_2O_3, Al_2O_3, MgO – соединения, вступающие в побочные реакции с H_2SO_4 и H_3PO_4.

Так для образования однородного крупнокристаллического осадка гипса необходимо, чтобы мольное отношение SO_3:CaO в жидкой фазе было в пределах 1,5-4,0 (1,5-2,5 % SO_3). Для полугидрата оно должно быть близко к стехиометрическому (0,8 – 1,2 % SO_3) [6]. Требования контроля и точности согласно технологическом регламенту АО «ФосАгро-Череповец» приведены в таблице 2.6.

Таблица 2.6. Основные показатели и требования точности аналитического контроля ЭФК.

Показатель	**Концентрационный диапазон, масс. %**	**Требуемая точность, абс. %**
SO_3	0,7 – 6,0	0,1 – 0,2
P_2O_5	25 – 55	0,5
H_2SiF_6	10,0 – 40,0	0,5
Fe	0,025 – 1,5	0,01
Осадок	< 5	0,2
Соотношение жидкость – твердое в пульпе	1,7 – 2,5 : 1	0,1

Таким образом ЭФК является ключевым полупродуктом (промежуточный объект) всей промышленной схемы и подлежит строгому аналитическому контролю. Кислота содержит 25 - 55 масс. % P_2O_5 и загрязнена примесями (Si, Al, S, Ca и Fe), содержание которых зависит от состава исходного сырья и типа производства. Косвенно наличие тех или иных примесей в ЭФК можно отследить и по химическому составу получаемых минеральных удобрений.

Далее, ЭФК нейтрализуется аммиаком для получения моноаммоний - (МАФ) или диаммоний - фосфатов (ДАФ), на основании которых получают сложные минеральные удобрения [2,4]. В отечественной практике реализована схема совместного получения ЭФК и сложных минеральных удобрений и используется схема получения МАФ и ДАФ и их последующего смешения с другими солями, содержащими питательные вещества [6,7]. Типовая схема производства сложных фосфорсодержащих минеральных удобрений приведена на рисунке 2.3 [2]. Данная схема адаптирована нами для общего случая.

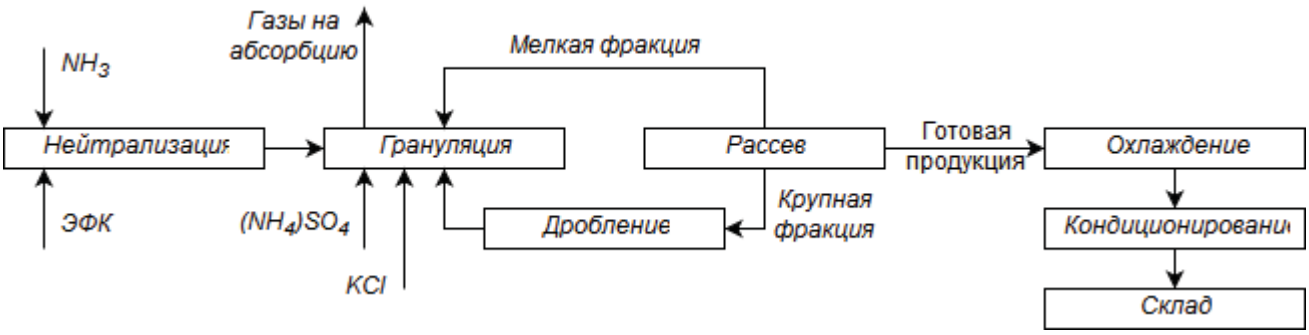


Рисунок 2.3. Общая схема производства сложных фосфорсодержащих минеральных удобрений.

В результате получается широкий спектр готового (товарного) продукта – минеральных удобрений, которые делятся на типы (PKS, NPK, NPS и.т.д.), характеризующиеся различным соотношением питательных и микро-элементов. Типы в свою очередь подразделяются на различные марки, характеризующиеся количеством тех или иных питательных элементов в агрохимикатах (таблица 2.7) [25–28].

При этом типы и марки сложных удобрений, а так же требования к точности анализа могут гибко изменяться в зависимости от требований сельского хозяйства, заказчиков и экономической ситуации [29]. Однако неизменными остаются требования аналитического контроля основных питательных элементов и микрокомпонентов (B, Mg, Ca, Zn), а так же показателей технологического процесса (гранулометрический состав, степень обработки кондиционирующими добавками и содержания H_2O).

На наш взгляд можно упростить подход и разбить весь спектр марок удобрений по матрице и диапазону определяемых концентраций. Так, каждый из исследуемых объектов (кроме апатита) можно представить смесями, максимально приближенными к технологическому процессу (таблица 2.8).

Таблица 2.7. Основные типы и марки удобрений производимые на предприятиях ФосАгро.

Тип удобрения	**Факторы, подлежащие контролю**	**Диапазон, масс. %**	**Требования точности, отн. %**
NPK марок 12-30-12, 12-32-12, 12-32-16, 6-20-30, 8-24-24, 16-16-8, 13-19-19, 9-20-20, 10-26-26, 9-25-25, 15-15-15, 10-20-10, 12-20-18, 16-16-16, 13-13-21, 7-30-20, 14-23-14	N	6 – 16	1
	P_2O_5	13 – 32	1
	K_2O	8 – 30	1
	S	3 – 11	0,5
	H_2O	< 1,0 – 1,5	0,15
	грансостав 1 – 6 мм	> 97	1
NP+S марок 20:20+14, 16:20+12, 19:38+7, 14:34+8	N	14 – 20	1
	P_2O_5	20 – 38	1
	S	7 – 14	1
	H_2O	< 1,0 – 1,3	0,5
	грансостав 2 – 5 мм	> 90	1
Аммофос (МАФ)	N	12 – 13	1
	P_2O_5	> 52	1
	H_2O	< 1,5	0,15
	грансостав 1 – 6 мм	> 97	1
ДАФ	N	18	1
	P_2O_5	47	1
	H_2O	< 1.8	0,15
	грансостав 1 – 6 мм	> 97	1

Таблица 2.8. Схема представления каждого объекта как искусственной смеси реактивов.

Тип объекта	**$NH_4H_2PO_4$**	**$(NH_4)_2SO_4$**	**$CaSO_4 \cdot 2H_2O$**	**K_2SO_4**	**S**
NPK	+	+	-	+	-
NPS	+	+	+	-	+
NPKS	+	+	+	+	+

Обоснованность данного подхода заложена в основе производства сложных удобрений и подтверждается в работах [2,3,5]. Исходя из первой части литературного обзора требуется провести полный, экспрессный, многоэлементный и динамический контроль основных продуктов производства: различные типы и марки сложных минеральных удобрений. При этом требуется предложить подход

подлежащий легкой автоматизации и комплексно описывающий весь процесс производства.

Заводские требования аналитического контроля производственных процессов включают в себя: проведение экспрессного (менее 15 минут) многофакторного и информативного анализа; анализ объектов для определения широкого спектра элементов (азот, сера, фосфор, примесные микроэлементы: железо, РЗЭ, тяжелые металлы); возможность проведения анализа в цеховых лабораториях.

Таким образом, нужен метод одновременного многоэлементного анализа с широким диапазоном определяемых параметров. Более того, метод должен быть автоматизирован и, при необходимости, встраиваться в производственную линию, работая в условиях цеха и цеховой лаборатории (повышенная пылимость, вибрации и нестабильность температуры), выполняться за 15 минут с момента отбора пробы до получения результатов.

2.3 Современные способы обработки аналитической информации

Ключевой особенностью аналитического контроля на промышленном производстве является постоянное стремление к автоматизации, сокращению влияния человеческого фактора, хранение получаемой информации – внедрению баз данных и комплексных подходов к контролю качества.

Повсеместное распространение получают лабораторные информационные системы контроля качества (ЛИМС), автоматические системы аналитического контроля (АСАК) и инструментальные методы, работающие в режиме «on-line» [30–33], обеспечивая экспрессность и полноту контроля. Максимальное время принятия решений для технолога предприятия составляет порядка 10-30 минут [34], именно за такой промежуток времени конечный продукт не успевает значимо измениться, а предприятие потерять значимое количество прибыли. На рисунке 2.4 представлена схема АСАК, работающего на апатит-нефелиновой обогатительной фабрике «ФосАгро» в г. Кировск.

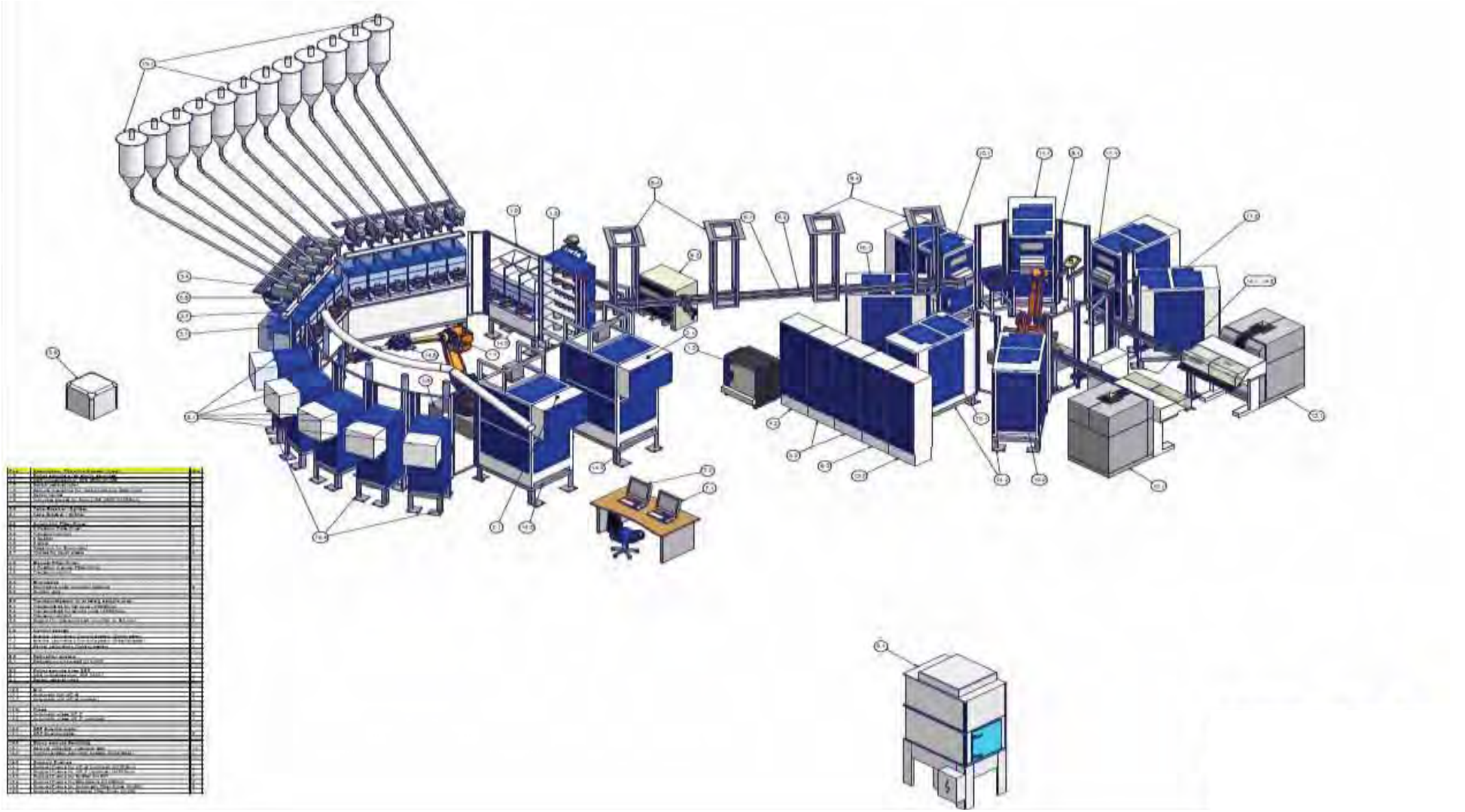


Рисунок 2.4. Схематическое изображение АСАК, работающего на апатит-нефелиновой обогатительной фабрике АО «ФосАгро» в г. Кировске.

Системы подобного типа и современные инструментальные методы аналитического контроля позволяют получать результаты анализов каждые 15-40 минут и накапливать огромные массивы информации. Однако данные установки дороги и требуют специального обслуживания, а получаемая информация требует создание баз данных и соответствующей статистической обработки [35]. Именно этим и занимается современное программирование и наука анализа больших массивов данных [31,36–50].

Анализ больших данных (АБД) – это комплексная, многомерная и непрерывная статистическая обработка постоянно увеличивающихся объемов характеристической информации [38], для создания автоматизированных программ, алгоритмов и баз данных для объективного, своевременного и комплексного контроля [39,50].

На сегодняшний день анализ больших данных позволяет решать следующие проблемы:

- выявление неявных зависимостей в данных, например расчета марки удобрения по азоту, прямое определение которого методами энергодисперсионного рентгенофлуоресцентного анализа (ЭД РФА) невозможно;
- сокращение количества градуировочных образцов и упрощение расчетов метрологических характеристик (таких как доверительный интервал) используя технологию «бутстрап» [40,42,48];
- использование автоматической классификации и визуализации исследуемых объектов, что повышает интерпретируемость данных и аналитических выводов [31,37–39,46,47,49,51];
- использование различных технологий классификации и регрессии, таких как логистическая регрессия, случайный лес, регрессия с регуляризацией и т.д. для предсказания физических и химических свойств исследуемых объектов - их концентраций и аналитических сигналов [31,36,41,45,52];

- использовать автоматический поиск и расчет наилучших уравнений связи аналитического сигнала с учетом множественных параметров [36,40,41,52].

И хотя данные подходы на сегодня применяются, в основном, в области квантовой и физической химии, медицине и реже в аналитической химии (достаточно в сжатой форме), они все больше и больше внедряются в смежные области. Описанные подходы созданы для анализа, визуализации и интерпретируемости. При этом описанные проблемы, которые решаются АБД явно выражены при управлении большим химическим производством (порядка 100 тонн в час на одну систему производства готового продукта, где буквально каждый грамм должен соответствовать требованиям нормативной документации и желаниям заказчиков).

Суть рассмотренного подхода заключается в создании базы данных на основе матрицы «объекты-признаки» [39,41,47], которая в дальнейшем обрабатывается с использованием следующих статистических методов:

- регрессия;
- классификация;
- понижение размерности данных и кластеризация.

Предварительно используется предобработка данных, такая как бутстрап, выравнивание выборок и нормализация данных. Технология бутстрапа представляет из себя непараметрический метод случайного выбора значений из общей выборки с возвращением, для формирования набора под выборок заданной величины. Описаны два подхода для аналитической химии [40,42,48]:

1. выбор объектов (т.е. пар значений «аналитический сигнал : концентрация»);
2. взятие среднего значения и добавление к нему случайного значения остатков (отклонений от среднего), представленных в выборке.

И хотя в литературе используются оба подхода, мы склонны использовать первый метод. Поскольку он признан классическим и обеспечивает наилучшее представление генеральной совокупности, причем не явно зависит от параметров распределения под выборки и может включать в себя скрытые взаимосвязи между

свойствами объекта. Остальные методы редко используются в аналитической химии и будут подробнее рассмотрены в теоретической части.

Регрессия является классическим методом поиска взаимосвязей в данных. Это, наверное, самый широко представленный в аналитической химии тип расчетов из рассматриваемых в АБД, однако зачастую достаточно неустойчивый. При этом требования к соотношению количества связанных признаков и известных объектов являются достаточно жесткими (порядка 1 к 5) [52,53] во избежание «переобучения» – когда математическая модель слишком сложна и идеально описывает тестовые данные, однако совершенно непригодна для анализа реальных объектов (рисунок 2.5).

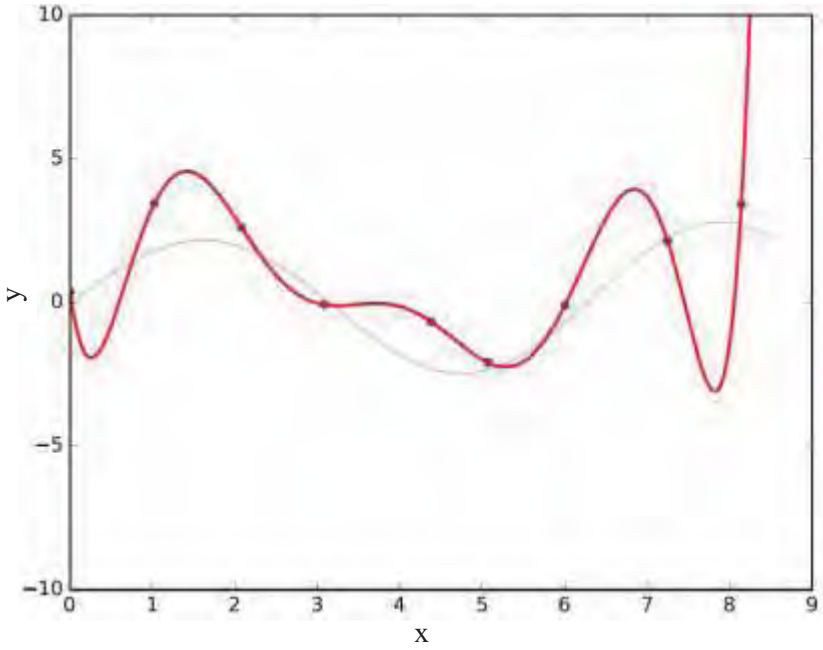


Рисунок 2.5. Иллюстрация переобучения на тестовых данных при слишком высокой степени регрессионного полинома.

Требования можно смягчить, грамотно отбирая признаки для модели – для этого и используют процедуру регуляризации, которая будет рассмотрена в теоретической части.

Регрессия, хоть и отмечается авторами как самый точный метод АБД из рассмотренных нами [36], чрезвычайно требовательна к исходным данным [52]. Так, данные должны быть сбалансированы, нормализованы и выравнены. Полученные результаты позволяют создавать многомерные динамические модели взаимосвязей, проводить комплексную регрессию и аппроксимацию, а так же

предсказывать неизвестный результат с заданной степенью достоверности [39–42,48,52].

Различные алгоритмы классификации заключаются в выборе оптимального способа отнесения объекта к тому или иному классу. В частности, широко используется технология «случайного леса», которая заключается в построении диаграмм выбора (да/нет) – так называемых «деревьев». При этом выбор производиться по набору числовых и логических параметров (значение интенсивности, наличие сигнала, порог значения и т.д.). Данный метод достаточно прост в расчетах и отлично подходит для выявления сложных и не явных зависимостей в данных [37,41,47,49,51]. Сегодня этот метод широко используется в медицине, экологии и всевозможной аналитике данных.

Исходя из проанализированной информации, становиться очевидным, что существуют пути создания автоматического или полуавтоматического комплекса по накоплению и обработке различных аналитических сигналов и физических свойств исследуемых объектов. И хотя подобные подходы пока еще не получили широкого распространения в аналитической химии и на химических производствах, проведенный обзор литературных данных отчетливо показывает преимущества анализа больших данных для аналитики в целом.

Для примера, нами была проведена работа по сравнению классического способа установления степени однородности стандартного образца и способа, с использованием сканирующего РФА с последующим АБД. В результате, была отмечена огромная информативность сканирующих методов рентгеновского анализа с АБД, в том числе и для выбора элементов, лимитирующих однородность образца [54] (рисунок 2.6).

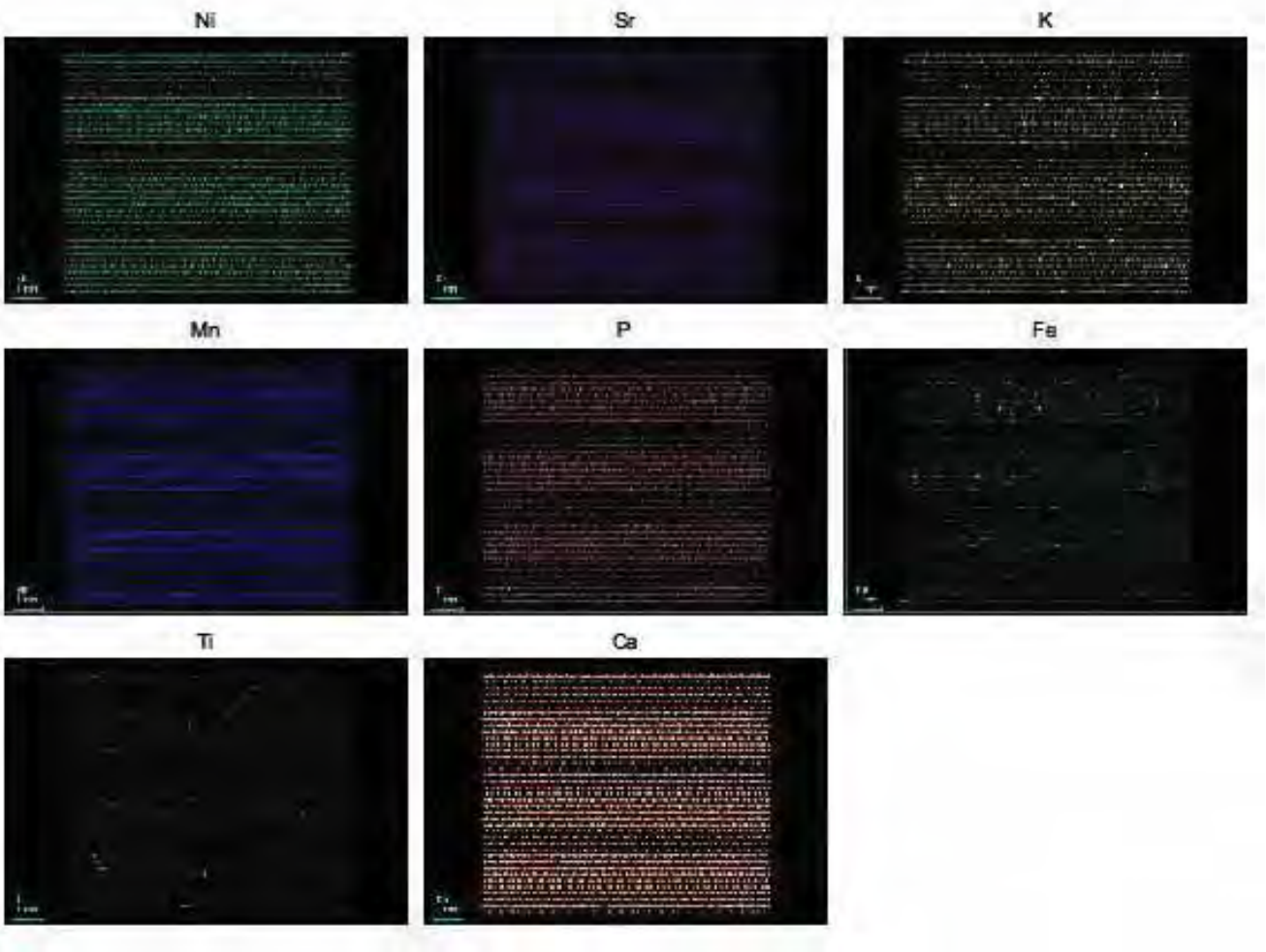


Рисунок 2.6. Визуальное распределение обнаруженных элементов в первой исследуемой пробе. Наличие специфических горизонтальных затемненных «полос» вызвано сжатием изображения

Даже визуально некоторые элементы имеют более равномерное распределение (например, кальций и фосфор) чем другие (железо, титан) на основании чего можно выбрать «элементы-маркеры», лимитирующие неоднородность. То же подтверждается и классическим способом [55] (таблица 2.9).

Таблица 2.9. Значение характеристик однородности распределения элементов в материале стандартного образца апатитовый концентрат (стандартные методики)

Элементы, тип пробы	**$F_{эксп.}$**	**$F_{табл}$, p=0,05**	**Вывод об однородности**	**Погрешность от неоднородности, %**
Ca, табл.*	2,60	5,96	Однородно	0,03
P, табл.*	2,29		Однородно	0,04
Fe, табл.*	4,03		Однородно	0,49
Ti, табл.*	0,21		Однородно	0,05
Ca, пор-к**	5,64	2,04	Не однородно	0,51
P, пор-к**	4,93		Не однородно	0,49
Fe, пор-к**	7,40		Не однородно	0,82
Ti, пор-к**	6,41		Не однородно	0,74

* таблетированная проба

** порошковая проба

Современная нормативная документация по разработке методик химического анализа [55–61] опирается на проверенные временем классические подходы к статистической обработке и представлению информации. Данные подходы заключаются к отборе качественной и представительной небольшой пробы из массы объекта [62] и проведения большого числа параллельный измерений [61]. Описанные подходы проверены и хорошо работают в рамках аналитической лаборатории, но когда речь заходит о непрерывном контроле современного большого химического производства и об отборе аналитической пробы весом 300 грамм из массы порядка 100 тонн – представительность и информативность таких подходов на наш взгляд будет сомнительна. Гораздо более правильным, интерпретируемым и информативным для технологии является статистический анализ больших данных. Например непрерывный контроль готовой продукции на транспортерной ленте (в режиме «on-line»), который хоть и не работает с тщательно подготовленной аналитической пробой, зато накапливает гораздо больше информации, лучше и точнее характеризующей производственный процесс [33,34,63,64].

То же относиться и к нивелированию влияния человеческого фактора и интерпретированию значимости погрешности. Ошибки при пробоподготовке и анализе гораздо более значимы для предприятия, когда одна точечная проба

представляет десятки тонн готовой продукции. С другой стороны, когда все условия зафиксированы и поступление информации непрерывно, данные ошибки перекрываются полноразмерной статистической обработкой больших массивов данных на большом объеме объектов.

Подобная картина наблюдается и при аттестации методик, когда повторяемость и воспроизводимость оцениваются по выбранному аналитическому сигналу в конкретном концентрационном поддиапазоне [61]. Тогда как машинное обучение позволяет динамически оценивать погрешность и автоматически вводить соответствующие поправки на всем протяжении общего диапазона концентраций [40,42,48], а так же наглядно и своевременно выводить полученную информацию о процессе.

Описанные подходы активно используются в компьютерной технологии и подлежат простой автоматизации, обеспечивая объективный, динамичный и понятный контроль различных процессов [39,46,50], пока что, к сожалению, не производственных.

Подводя промежуточный итог современным способам работы с информацией, становятся очевидными пути применения изученных подходов в промышленной практике. Поставленные задачи предлагается реализовать не как систему автоматического контроля, подключенную к информационной среде предприятия, а в едином программном и аппаратном комплексе для учета и изучения различных взаимосвязей между физическими свойствами исследуемых объектов и их химическим составом. Описанные подходы позволят создать экономически выгодный и эффективный метод комплексного контроля производства с организацией единой, постоянно развивающейся базы данных «объекты-признаки». Таким образом закрывается ниша полуавтоматического контроля, обладающего информативностью и комплексностью АСАК, а также простотой реализацией классических методов. Для данных целей остается выбрать информативный аналитический способ получения информации о химическом составе и физических свойствах исследуемых объектов.

2.4 Методы аналитического контроля производства минеральных удобрений

Исходя из особенностей основных исследуемых объектов нами были выделены их свойства и элементы-маркеры качества (таблицы 2.4, 2.6 и 2.7), требования к скорости анализа (а значит и к пробоподготовке) и основные математические методы работы с аналитическими сигналами (а значит и требования к информативности получаемых аналитических сигналов). Но какие же существуют физические и химические методы для исследования рассмотренных объектов?

По одной из принятых классификаций аналитические методы можно разделить на молекулярные и атомарные [65,66]. Молекулярные – определяют ту или иную форму химического соединения, в свою очередь атомарные – определяют содержание тех или иных элементов в образце, вне зависимости от их молекулярного состояния.

Стоит отметить, что для контроля основных питательных элементов часто не требуется знать их молекулярную форму, ведь процесс производства отлажен и получаемые соединения известны. Именно по этому при производстве приняты особенные формы пересчета для каждого из определяемых элементов (таблица 2.10) [2–4,6,7,67,68].

Таблица 2.10. Определяемые элементы и их формы пересчета

Элементы	**Формы пересчета**
P	P_2O_5
S	S, SO^-_3, SO^-_4
Cl	Cl^-
Si, Na, K, Ca, Mg, Zn	SiO_2, Na_2O, K_2O, CaO, MgO, ZnO
Fe	Fe_2O_3

При этом форма пересчета не зависит от реального содержания той или иной молекулярной формы. Таким образом нам не важно, является ли наш метод атомарным или молекулярным.

Проанализировав различные группы методов аналитического контроля, используемые для анализа выбранных нами элементов в различных матрицах [34,63,64,68–86], можно составить сводную таблицу для групп методов (таблица 2.11):

- энергодисперсионный рентгенофлуоресцентный анализ (ЭД РФА);
- РФА с полным отражением (ПО РФА);
- волнодисперсионный (ВД) РФА;
- масс-спектрометрия с индуктивно связанной плазмой (ИСП МС);
- атомно-эмиссионная спектрометрия (АЭС) с ИСП;
- атомно-абсорбционная спектрометрия (ААС) с ИСП;
- ААС;
- АЭС;
- титрование;
- гравиметрия.

Таблица 2.11. Группы методов аналитического контроля

Тип метода	**Чувствительность**	**Селективность**	**Многоэлементость**	**Возможность автоматизации**	**Длительность анализа*, мин**	**Длительность пробоподготовки*, час**
ЭД РФА	Средняя	Средняя	+	Высокая	0,2 - 10	0 - 3
ПО РФА	Высокая	Средняя	+	Низкая	0,2 - 10	0 - 3
ВД РФА	Высокая	Высокая	+	Высокая	0,2 - 10	0 - 3
ИСП-МС	Высокая	Средняя	+	Низкая	1	0,5 - 3
ИСП-АЭС	Высокая	Высокая	-	Средняя	1	0,5 - 3
ИСП-ААС	Высокая	Высокая	-	Средняя	1	0,5 - 3
ААС	Высокая	Высокая	-	Средняя	1	0,5 - 3
АЭС	Средняя	Средняя	-	Средняя	1	0,5 - 3
Титрование	Высокая	Средняя	-	Низкая	от 10	0,5 - 9
Гравиметрия	Высокая	Высокая	-	Низкая	от 30	1 - 9

* указана приблизительно – для объектов типа горные породы, почвы, минеральные удобрения (одна проба).

В современной отечественной промышленности, как правило, используются классические методы анализа, такие как спектроскопия, титрометрия, гравиметрия, реже атомарные: атомно-абсорбционная спектроскопия. Это вызвано тем, что до недавнего времени задачей выходного контроля было только определение основных питательных веществ и сложно было представить комплексный анализ не только всего производства в целом, но даже и конкретного объекта. Дополнительным фактором являлась доступность в плане цены и квалификации персонала.

Однако на сегодняшний день встает задача определения не только конкретных элементов в матрице, но и поиск взаимосвязей содержания тех или иных элементов с протеканием всего технологического процесса, а так же непрерывный количественный контроль последнего. Например описанная авторами [87] схема управления производством фосфорной кислоты базируется на таких аналитических показателях, как:

- концентрация серы в жидкой фазе пульпы;
- концентрация фосфора в жидкой фазе пульпы;
- концентрация твердой фазы в пульпе;
- уровень и температура пульпы.

Данные показатели предсказываются и регулируются не только на основании аналитического контроля, но и на основании любых доступных показателей, таких как расход потоков реагентов, плотность, цветность и т.д.

Подобные работы по взаимному учету множества доступных параметров и поиску их связей с производственным процессом ведутся и другими отечественными и зарубежными исследователями [33,34,67,68,82]. Становиться очевидным что промышленность интересует не столько одна конкретная цифра, сколько полнота и селективность извлечения информации из поточного и автоматического анализа. Наиболее информативными методами являются спектральные методы, представляющие общий набор информации об объекте. В

особенности методы, основанные на общих физических и химических свойствах объектов, таких как масс-спектрометрия и РФА.

Это подтверждается и результатами поиска в системе «Scopus» [88]. Для выделенных типов аналитических анализов можно проследить следующую динамику количества научных работ за последние 10 лет (рисунок 2.7).

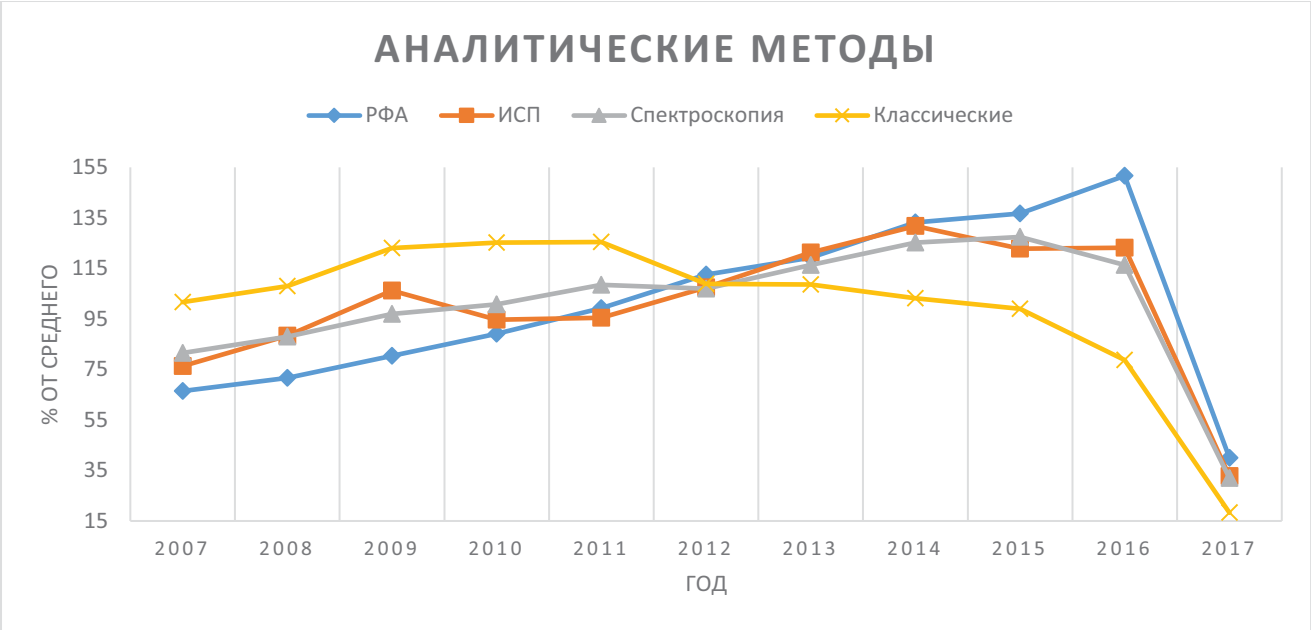


Рисунок 2.7. Динамика количества статей, посвященных тому или иному аналитическому методу, нормированных на среднее количество статей в год. Выборка состоит из статей, опубликованных за последние 10 лет по результатам «Scopus». В «РФА» объединены ЭД и волновой методы; «ИСП» - различные виды спектроскопии с индуктивно связанной плазмой (ИСА); «Спектроскопия» - методы спектроскопии без использования ИСП; «Классические» - титрование и гравиметрия.

Из обзора методов можно заметить, что РФА более динамично развиваются в аналитической области. При этом, если выделить из всего массива статьи, относящиеся к теме окружающей среды и агрохимии (таблица 2.12), то к финишу выходят два фаворита: РФА и ИСП методы. И это не удивительно, ведь данные методы позволяют получать информацию о веществе «напрямую» - не используя длительную пробоподготовку (производные формы, дополнительные химические реакции и т.д.). Так же данные направления имеют возможность многоэлементного анализа, что обеспечивает высокую информативность результатов.

Таблица 2.12. Процент статей от общего числа за 10 рассмотренных лет по тематике «Environmental Science», «Earth and Planetary Sciences» и «Agricultural and Biological Sciences»

Аналитический метод*	**Процент статей от общего числа, %**
РФА	18,57
ИСП	21,50
Спектроскопия	5,81
Классические методы	7,42

* В «РФА» объединены ЭД и волновой методы; «ИСП» - различные виды спектроскопии с индуктивно связанной плазмой (ИСА); «Спектроскопия» - методы спектроскопии без использования ИСП; «Классические методы» - титрование и гравиметрия.

При этом, если индуктивно-связанная плазма с масс-, атомно-абсорбционной или эмиссионной спектроскопией чувствительна к агрегатному состоянию объекта и концентрации элементов (позволяет работать только с разбавленными жидкостями) и требует дополнительных реагентов [64,69,73,76,78,81,85,89], то рентгеновские методы лишены данных недостатков и позволяют комплексно решить поставленные в работе задачи.

Это же отмечается многими авторами, работающими с данным классом методов. Так в ряде работ [32,75,79,90,91] приводиться комплексный анализ различных удобрений с использование РФА, а в нормативной документации [92] описывается использование единственного метода – РФА, для полного аналитического контроля фтористого алюминия, одного из побочных продуктов при переработке апатита. В ряде работ приводиться сравнение аналитических и статистических характеристик описанных методов (таблица 2.13).

При этом большое количество авторов делает акцент на возможность онлайн, поточного и автоматизированного контроля с использованием РФ-методов [32–34,64,68,70,93–96]. Так же встречаются работы где отмечается огромная информативность РФА и используются подходы анализа больших данных: понижение размерности и визуализация, множественный поиск корреляций, множественная регрессия с учетом различных аналитических и физических показателей [68,76,77,90,97].

Таблица 2.13. Сравнение заявленных пределов обнаружения и СКО для разных объектов и методов

Метод	**Объект**	**Аналиты**	**Определяемые концентрации, %**	**Относительное отклонение от среднего или проверочного метода, %**
ИСП-АЭС [76]	удобрения	Fe_2O_3 MgO MnO Na_2O_3 P2O5	0,79 0,33 0,017 0,86 30,2	4 4 3 3 3
ИСП-МС [79,98]	удобрения	Ca K Mg P Zn B Si	-	3,47 2,93 3,72 2,74 1,17 2,2 3,82
ЭД РФА	почвы [99]	K Ca Ti Fe Ni Cu	2 1 0,5 3,5 0,002 0,003	20,65 7,42 1,92 0,84 4,88 25
	сточные воды [71]	Mg Si P S K Ca Fe Zn	<1 <3 <10 <10 <50 <30 <6 <1	2,10 0,49 0,80 0,48 0,24 0,09 2,93 2,78
	не фосфатные минералы [100]	K Ca Fe Zn	1,21 16,3 2,57 0,002	2 - 20 6 - 16 1 - 10 2 - 7
ВД РФА	фосфатные минералы [67,68,81]	MgO Na_2O Al_2O_3 SiO_2 P_2O_5 SO_3 Cl CaO SiO_2 Fe_2O_3 MnO	0,3 - 1,6 < 1 0,2 - 1,8 4,1 - 11,2 15,0 - 37,0 < 1 < 1 43,6 - 56,3 4,1 - 11,2 0,2 - 1,1 0,01 - 0,03	0,01 - 4,5 3,5 0,04 - 9,3 2 0,22 - 0,44 1,5 6,5 0,37 - 0,5 0,2 - 0,4 0,03 - 0,06 0,001-0,002

Проведя литературный обзор аналитических методов, становиться очевидно, что для решения поставленных задач в промышленных условиях и полного информативного описания исследуемого объекта метод РФА является практически безапелляционной альтернативой по цене, информативности и приборной реализации.

При этом существует несколько разновидностей РФА:

- энергодисперсионный;
- волнодисперсионный;
- полного отражения.

Некоторые исследователи выделяют еще и портативный (переносной) вариант РФА, однако с точки зрения аппаратной части, способов пробоподготовки и обработки информации он мало отличается от ЭД РФА и в нашей работе отдельно рассматриваться не будет. Каждый из описанных подходов обладает своими сильными и слабыми сторонами. Они зависят от типа объекта, пробоподготовки и условия анализа. Основные характеристики РФА обобщены и приведены в таблице 2.14 [32,33,63,67,70,74,75,77,90,91,93,99,101–105].

Исходя из условий комплексного получения информации, экспрессности определения широкого набора элементов и возможности автоматизации нами выбран энергодисперсионный РФА как наиболее неприхотливый к условиям, информативный и гибкий метод анализа. Авторы изученных работ сходятся во мнении, что ЭД-РФА является достаточно чувствительным и селективным методом для работы с исследуемыми объектами.

По результатам проведенного обзора, метод ЭД-РФА удовлетворяет нашим требованиям для решения поставленных задач и будет в дальнейшем использоваться в представленной работе. Метод применим в комплексном подходе к анализу: поиску скрытых взаимосвязей между физическими параметрами (такими как pH, растворимость и т.д.) и РФ спектром, использованию множественной регрессии и АБД [90,99,105–107].

Таблица 2.14. Основные характеристики ЭД, ВД и ПО РФА для объектов типа «горные породы», «удобрения»

	ЭД РФА	**ВД РФА**	**ПО РФА**
Чувствительность	C - F от 1 % Na - S 0,01 % Cl - U 0,0001 %	C - F от 0,01 % Na - S 0,001 % Cl - U 0,00001 %	C - F от 0,01 % Na - S 0,001 % Cl - U 0,0001 %
Перекрывание линий спектра	Высокая	Низкая	Средняя
Диапазон определяемых элементов	C - U	Зависит от установленного кристалла, но возможно от Be до U	C - U
СКО от среднего или установленного значения*, %	< 21	< 6	< 10
Время получения информации о всем спектре (скорость сканирования спектра)	> 5 с	Зависит от схемы: > 25 с для многоканальных, но информация об ограниченном наборе элементов > 10 мин для сканирующих (низкая)	> 5 с
Разрешающая способность	Низкая (100 - 300 эВ)	Очень высокая (5 - 20 эВ)	Низкая (100 - 300 эВ)
Мощность, Вт	< 50	< 4000	< 50
Эффективность возбуждения	Низкая	Высокая	Низкая
Фоновая составляющая	Высокая	Низкая	Низкая
Прочие требования	Легкое встраивание в производственные помещения и использование дополнительных методов получения информации (фото- и видеосъемка, преконцентрирование и т.д.). Пониженные требования к вибрации, пыли, влажности и т.д.. Возможность компактного исполнения.	Повышенные требования в внешним условиям (вибрация, пыль, влажность и т.д.). Высокая стоимость.	Основан на ЭД схеме с измененными углами падения и выхода излучения. Обладает более высокими требованиями к поверхности образца и пробоподготовке в целом

2.5 Энергодисперсионный метод рентгенофлуоресцентного анализа

Выбрав метод, остается рассмотреть особенности его применения и те возможности и проблемы, которые скрываются за термином ЭД-РФА. Метод обладает известными положительными свойствами:

- минимальная пробоподготовка [63,90,108,109]:
  - перетирание и прессование порошков;
  - прямой анализ жидкостей и твердых тел;
- многоэлементность и большое количество получаемой информации (спектр по всем элементам от углерода до урана с учетом плотности образца) [32,99,101,105];
- гибкость приборной реализации (от стационарных до переносных приборов) [63,78];
- пониженные требования к условиям окружающей среды (наличия вибраций, пыли и повышенной влажности) [106,108,109];
- легкая возможность автоматизации и установки в производственных условиях [32,33,82,110].

И набором недостатков:

- низкая чувствительность [78,99,110];
- нехватка разрешающей способности [105];
- матричное влияние [105,111];
- отсутствие методик контроля качества получаемой информации [105].

Так, автор работы [105] отмечает, что из-за описанных недостатков ЭД РФА часто применяется в качестве полуколичественного «сканирующего» контроля на различных производственных линиях (при производстве полупроводников, добычи сырья, геологических работах и т.д.).

Однако в случае анализа выбранных нами объектов исследования наиболее значимыми недостатками являются матричные эффекты и сложность предсказания качества получаемой информации:

- неоднородности пробы и облучаемой поверхности;
- влажность и непостоянная плотность продуктов;
- возможность изменения элементного состава;
- требование ускоренной пробоподготовки.

Отдельным немаловажным пунктом выступает требование комплексного подхода к анализу объектов, контроль:

- цветности,
- наличия посторонних включений,
- химического состава и т.д.;

Рассмотрим наиболее часто встречающиеся пути устранения мешающих влияний для подобного типа объектов и определимся со стратегией проведения нашей работы.

Как упоминалось ранее, требование экспрессности и информативности анализа является пожалуй ключевым критерием на производстве. Как правило требуется принимать решения в течении 15 минут, опираясь на аналитические данные о процессе. И минимальная пробоподготовка сохраняющая полноту физических свойств объекта (цветность, плотность, соотношение элементов) как нельзя лучше подходит для данных целей. Как правило выделяют несколько подходов к пробоподготовке [105], которые в нашем случае ограничены требованием сохранности исходных физических и химических свойств. Таким образом остается единственный вариант – прессование или измерение твердых объектов напрямую.

Однако даже в рамках этого подхода существует множество вариаций: подбор степени измельчения, режима сушки, использования разбавления и т.д., что так же принято называть матричными эффектами и учитывать с помощью различных математических и физических подходов (таблица 2.15).

Таблица 2.15. Основные способы пробоподготовки с сохранением физико-химических свойств исследуемых объектов

Объект	**Тип пробоподготовки**	**Дополнительные реактивы**	**Время приготовления одной пробы**	**Условия хранения при устойчивости во времени**
твердые	прямое измерение	нет	1 мин	на воздухе
	измельчение до порошка		10 мин	в эксикаторе
порошки	высушивание, прямое измерение	нет	1 мин без высушивания	в эксикаторе
	прессование	нет или различные связующие (борная кислота, ПВС* и т.д.)	5 мин	
	сплавление	флюсы (тетраборат лития и натрия и т.д.)	10 мин	на воздухе

* поливиниловый спирт

Следующей по величине проблемой ЭД РФА является учет матричных влияний. В литературе данное явления характеризуется [105,112]:

- крупность частиц порошка или однородность жидкости;
- плотность матрицы объекта:
  - фон (рассеянное излучение);
  - ширина пиков;
- толщина излучателя:
  - насыщенный слой;
  - тонкий излучатель;
- элементный состав:
  - до поглощение характеристической линии;
  - до возбуждение характеристической линии;
  - перекрывание пиков.

Каждый эффект присутствует в исследуемых объектах и достаточно хорошо изучен в литературе. Так, в работе [105] составлена информативная диаграмма способов учета матричного влияния (рисунок 2.8).

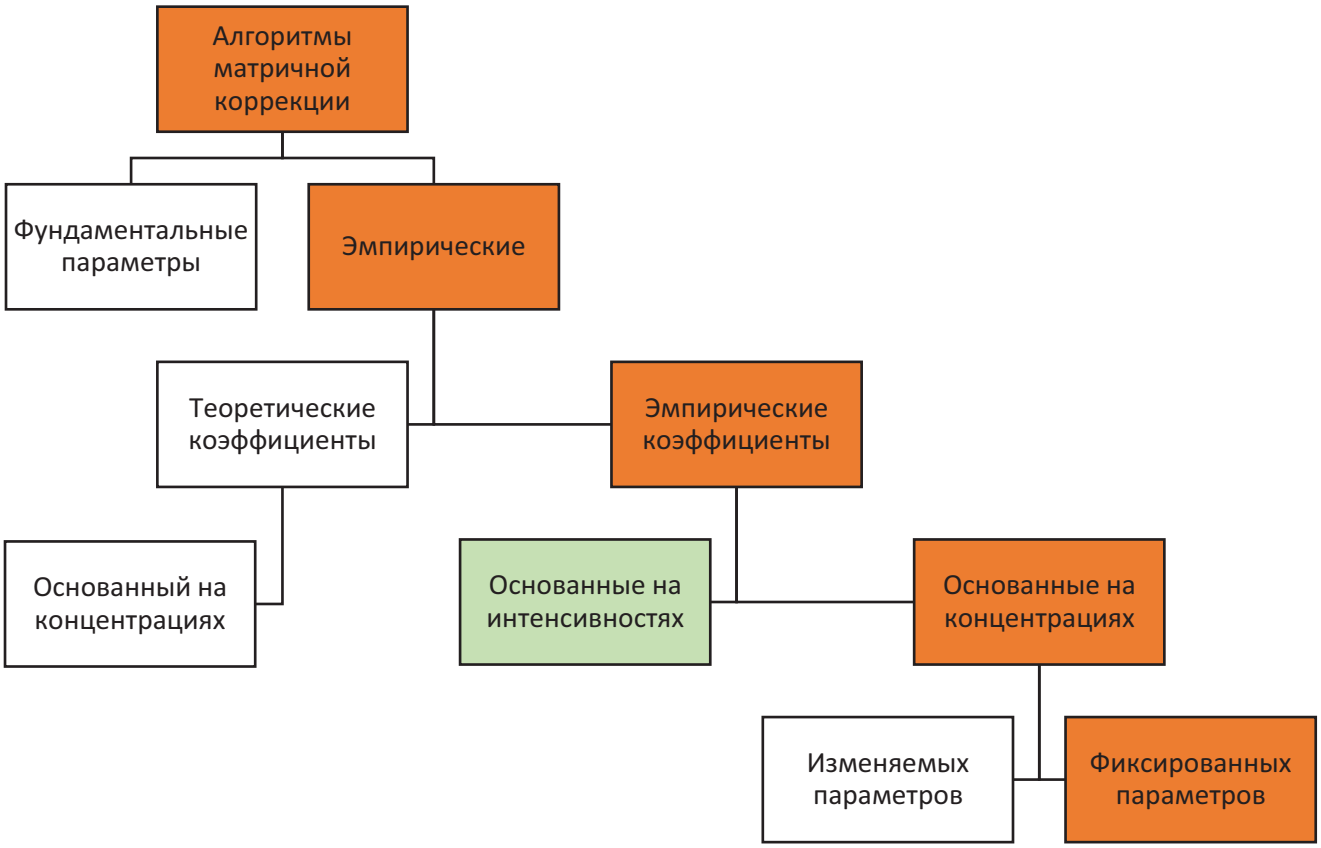


Рисунок 2.8. Общий алгоритм классического решения задачи учета матричных эффектов.

Путь выбранный автором [105] обозначен оранжевым цветом и базируется на методе Лукаса-Туса, однако в нашем случае множественного учета параметров наиболее оптимальным представляется путь основанный на эмпирических коэффициентах и работе с интенсивностями, один из вариантов «множественной» регрессии (отмечен зеленым цветом).

На основании проведенного обзора составлена таблица с наиболее распространенными способами учёта матричных эффектов (таблица 2.16). К сожалению, в найденной литературе авторы используют традиционные зависимости, основанные на явной физической природе спектра, такие как пик некогерентного рассеяния, учет влияния ближайших элементов или поглощения матрицей.

Однако не стоит забывать и о современных подходах АБД, которые позволяют комплексно учитывать и обобщать различные физические и математические параметры, а так же находить скрытые взаимосвязи между

признаками. Тогда каждый из рассмотренных матричных эффектов можно оценить по выделенным свойствам пробы, таким как интенсивность аналитического сигнала, крупность фракции и т.д.

Следующими возможными проблемами являются разрешающая способность и чувствительность анализа, которые решаются подбором условий измерения [53,86], работой с интенсивностями, вместо площадей пиков и различными математическими подходами к удалению фона и аппроксимации характеристических линий [105], которые подробнее будут обсуждаться в теоретической части данной работы.

Используя полученную информацию и разобранные ранее способы работы с большими данными, нами предлагается использовать всю информацию спектра для выявления неявных влияний на аналитический сигнал и создания универсальной и автономной модели для технологических объектов. Наиболее близко к этому подошли методы множественной регрессии [105,112–114], хотя авторы используют только простую регрессию без регуляризации с использованием случайно выбранных или явных (основанных на физических представлениях о спектре) регрессоров. В нашем же подходе предлагается использовать регрессию с регуляризацией по всему спектру, что позволит автоматически отбирать значимые для анализа факторы и подстраивать модель под исследуемый объект. Данный подход, основанный на создание единой базы данных «объект-признаки», позволить автоматизировать выбор уравнения связи и создать гибкое единое представление промышленных объектов на протяжении всего процесса производства.

Таблица 2.16. Матричные эффекты и способы их устранения

метод	**учитываемые влияния**	**математическое выражение**	**Количество используемых каналов**
линейная зависимость	нет	C=AX+B, C - концентрация X - аналитический сигнал	1 - 10
множественная регрессия	элементы матрицы, когерентное/некогерентное рассеяния (сама матрица), в теории любое (даже без физического смысла)	$C=f(X, J_i)$, C - концентрация X - аналитический сигнал Ji - регрессоры, учитывающие элементы матрицы и прочие факторы	$N(J_i)$ * 10 - 4000 (в теории - все каналы)
фундаментальные параметры	дополнительное поглощение и возбуждение характеристического излучения матрицей объекта	$C = f(X, C_i, F_i)$ C – концентрация аналита Ci – концентрации мешающих элементов Fi – фундаментальные параметры, зависящие от матрицы	1 - 10
эмпирические методы с теоретическими коэффициентами (например Брикса, Ланчастера-Трейли и т.д.)	аналогично фундаментальным параметрам с эмпирическими приближениями некоторых коэффициентов	$C = f(X, C_i, F_i)$ плотность матрицы (выраженная с использованием коэффициентов поглощения элементов или пиками рассеяния трубки)	1 - 10
эмпирические методы с эмпирическими коэффициентами (метод Руссо, стандарта-фона, Лукаса-Туса)	дополнительное поглощение и возбуждение характеристического излучения матрицей объекта	$C = f(X, N_i, a, u)$ C - концентрация X - аналитический сигнал; Ni - элементы матрицы, a – «допоглощение» от матрицы, u – «довозбуждение» от матрицы	(1 - 10) * N_i

Обобщая проведенную работу нам представляется определенный цикл работы с методами ЭД-РФА (рисунок 2.9)

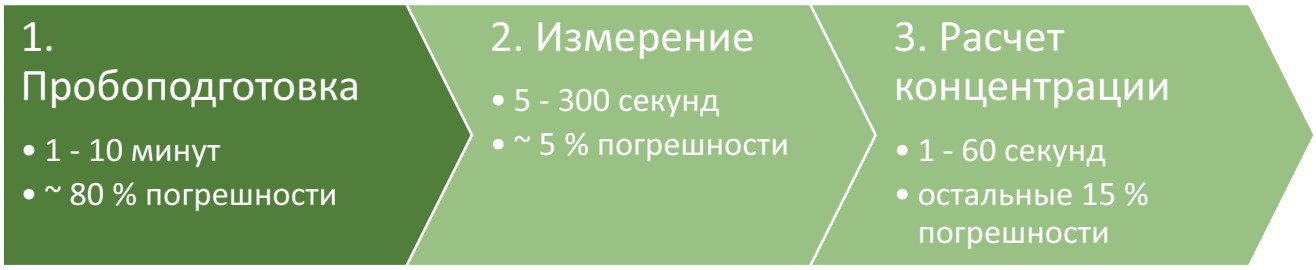


Рисунок 2.9. Общий цикл работы с ЭД-РФА.

Когда же методы расчета концентрации оказываются не эффективны или слишком сложны, часто прибегают к дополнительной пробоподготовке (возвращаются на 1 стадию цикла). Однако это в разы увеличивает время проведения и погрешность анализа, что совершенно неприемлемо в нашем случае.

Таким образом, наряду с формированием проб для анализа и подбором условий измерения, нами будут рассмотрены различные способы пересчета аналитического сигнала в концентрации с использованием всей информации заложенной в спектре, а так же любой информации, доступной для анализа. Предложенный подход на наш взгляд позволит реализовать точный, информативный и универсальный метод мониторинга промышленного процесса.

2.6 Заключение

Проведя обзор литературы и выделив необходимые инструменты стало ясно как нужно провести нашу работу и решить поставленные задачи.

Изначально нашими задачами являлись:

1. Проанализировать процесс промышленного производства сложных фосфорсодержащих удобрений и выявить ключевые объекты контроля.
2. Выделить значимые химические и физические параметры для эффективного учета сложной матрицы исследуемых объектов.
3. Разработать конкурентную и экономически эффективную систему экспрессного получения информации об объектах анализа.

4. Разработать автоматический алгоритм выделения и расчета информативных признаков при контроле качества производимой продукции.
5. Создать прототип единой аналитической базы исследуемых объектов с возможностью использования методов АБД.
6. Теоретически обосновать и разработать оптимальную схему комплексного анализа сложных фосфорсодержащих удобрений для обеспечения условий наиболее экспрессного перехода с одной производимой марки на другую.
7. Обеспечить простоту и удобство пользования данным методом в заводских лабораториях (упрощение программной, процедурной и аналитической части приборов).

Наш обоснованный и практически безапелляционный инструмент аналитического контроля – метод энергодисперсионного рентгенофлуоресцентного анализа и анализ больших данных, при которых каждая задача имеет свои пути решения:

1. Задачи физико-химического представления исследуемых объектов становиться возможно решать комплексно, используя информацию всего спектра (порядка 4000 каналов) и физических свойств объектов (крупности частиц, плотности поверхности и т.д.) и создав единую базу данных «объекты-признаки».
2. Динамический контроль, обеспечение стабильности и воспроизводимости решается АБД с использованием регрессии и классификации.
3. Описанные методы аналитического контроля и математической обработки легко автоматизируются и встраиваются в ПО и базы данных, что так же упрощает наши задачи.

Дополнительно требуется:

1. Разработать простой и экспрессный метод получения информации о физических свойствах пробы (например, фракционном составе) в дополнение к ЭД РФ методу получения информации о химическом составе пробы.
2. Выделить значимые химические и физические параметры для эффективного учета сложной матрицы исследуемых объектов.

3. Разработать конкурентную и экономически эффективную систему экспрессного получения информации об объектах анализа.
4. Разработать автоматический алгоритм выделения и расчета информативных признаков при контроле качества производимой продукции.
5. Обеспечить простоту и удобство пользования данным методом в заводских лабораториях: упростить программную, процедурную и аналитическую часть разрабатываемого подхода.

Результатом работы будет являться алгоритм внедрения РФ-метода в процесс производства минеральных удобрений, обеспечение автономности его работы, а также постоянное накопление и обработка количественной и качественной аналитической информации по всем доступным физико-химическим показателям процесса производства.

3 Теоретическая часть

В данной главе приводится математическая постановка решаемых задач, основные определения, формулы и алгоритмы, использованные в работе. Описываются пути оптимизации и визуализации данных. Все описанные подходы автоматизированы и реализованы автором на языке Python 2.7. По оптимизированным алгоритмам написана и зарегистрирована программа «DSpectra». Ключевые способы программной реализации описанных алгоритмов приведены в приложении А.

3.1 Определения величин и решаемая задача

В работе использованы следующие определения:

$\mathbb{X}$	пространство объектов
$\mathbf{y = y(x)}$	ответ на объекте
$\mathbb{Y}$	пространство ответов на объекте
$\mathbf{x = (x_1, x_2, ..., x_d)}$	объект, имеющий d-мерное признаковое описание
$\mathbf{D_j}$	множество значений j-ого признака по всем объектам
$\mathbf{X=(x_i, y_i)^l_{i=1}}$	обучающая выборка
a(x)	алгоритм предсказания (для классификации или регрессии)
$\mathbb{A}$	пространство возможных алгоритмов предсказания
Q(a, X)	функционал ошибки
гиперпараметры	настраиваемые параметры для оптимизации

В работе рассматриваются признаки трех типов:

бинарные	$D_j = \{0,1\}$
вещественные	$D_j = \mathbb{R}$
категориальные	D_j – неупорядоченное множество объектов, для которых не возможны операции сравнения

Задачей анализа больших данных ставиться подбор такого алгоритма *a(x)*, для которого достигается необходимый минимум функционала ошибки:

$$a(x) = argmin_{a \in \mathbb{A}} Q(a, X)$$

При этом функционалы ошибки могут отличаться в зависимости от решаемой задачи: классификации или регрессии и по своему математическому смыслу. Часто для описания работы алгоритмов вводят «метрики качества» - величины, прямо или косвенно связанные с функционалом ошибки и характеризующие предсказательную силу алгоритма.

3.2 Обоснование возможности усовершенствования предсказательной силы спектрометра

На сегодняшний день в аналитической химии используется лишь малая часть признаков исследуемых объектов для построения уравнений связи. Если в случае классических методов анализа, таких как титрование, гравиметрия или спектроскопия описанный подход является практически единственно возможным, то в случае ЭД РФА данный способ кажется более чем расточительным. По стандартному, не обработанному, РФ спектру, как правило, представлено 4096 каналов, в каждом из которых записана информация о количестве импульсов с определенной энергией. При этом каналы как правило линейно зависят от энергии излучения:

$$E = a_0 + a_1 \cdot N$$

где: *E* – энергия канала, a_0 и a_1 – коэффициенты пересчета, уникальные для каждого детектора, *N* – номер канала.

Даже с учетом обнуления некоторых каналов и работы спектрометра на режимах пониженного напряжения для плотных матриц (во избежание перегрузки детектора и, как следствие, увеличения мертвого времени) остается информация порядка 2000 каналов. С учетом взаимного влияния энергий и фундаментальных параметров выбор 1 – 20 каналов из 2000 для построения уравнений связи кажется нам необоснованной потерей информации. В работе предложены

автоматизированные механизмы расчета необходимого количества признаков из всего спектра без потери информации на основе методов машинного обучения и анализа больших данных.

Учитывая упомянутую сильную зависимость спектра от физических свойств пробы и требования технологического контроля к учету множества химических и физических свойств: содержание тех или иных элементов, гранулометрический состав, степень обработки кондиционирующими добавками и т.д., в дополнение к ЭД РФА предложено дополнение в виде системы оптического контроля.

Таким образом, теоретическая часть работы заключается в обосновании и разработке автоматизированного и комбинированного метода получения данных с оптико-ЭД РФ системы для дальнейшего анализа химических и физических свойств исследуемых объектов методами «машинного обучения» - автоматизированных алгоритмах классификации и регрессии на больших данных.

3.2.1 Выделение физических признаков пробы с использованием оптической системы анализа

Предложенная схема оптического анализатора позволяет работать с твердыми объектами в виде гранул, прессованных гранул и прессованного порошка различных фракций. Математический аппарат установки включает в себя следующие процедуры.

1. Возможность записи фотографии разрешением не менее 640×480 как трехмерную матрицу интенсивностей пикселей по системе RGB (Red Green Blue, Красный Зеленый и Желтый каналы соответственно).
2. Выделение участка фотографии с поверхностью пробы и расчет его площади в пикселях.
3. Форматирование исходной трехмерной матрицы в двухмерную матрицу яркости как среднее по трем каналам RGB каждого пикселя:

$$I = \frac{I_R + I_G + I_B}{3}$$

где: I, I_R, I_G, I_B – общая яркость и интенсивности по красному, зеленому и синему каналам соответственно.

4. Дифференцирование для устранения трендов освещенности и получения необработанной «карты поверхности» пробы – двумерной матрицы интенсивностей пикселей в градациях серого, характеризующей поверхность исследуемого объекта:

$$dY_i = Y_i - Y_{i-1}$$

5. Сглаживание карты поверхности медианным фильтром для устранения шумов:

$$y_i = median_{i=-w}^{w}(Y_i)$$

где: y_i – сглаженная яркость пикселя, Y_i –матрица яркости области i-ого пикселя, w – окно сглаживания. Медианный фильтр выбран как наиболее приемлемый для сохранения границ объектов на изображении [115].

6. Приведение карты поверхности к бинарному виду (бинаризация) по среднему порогу:

$$y_i = \begin{cases} 1, & y_i > k \\ 0, & y_i < k \end{cases}$$

$$k = average(y)$$

где: y_i – интенсивность яркости пикселя, k – константа – средняя интенсивность изображения (если не оговорено обратное).

7. Программная аппроксимация аномалий как эллипсов и расчет их удельного количества и удельной площади по алгоритму «совпадающих квадратов». Для данного алгоритма изображение разбивается на определенное количество областей (квадратов) по каждому из которых производиться поиск определенной структуры (рисунок 3.1). Более подробное обсуждение данного алгоритма выходит за рамки настоящей работы.

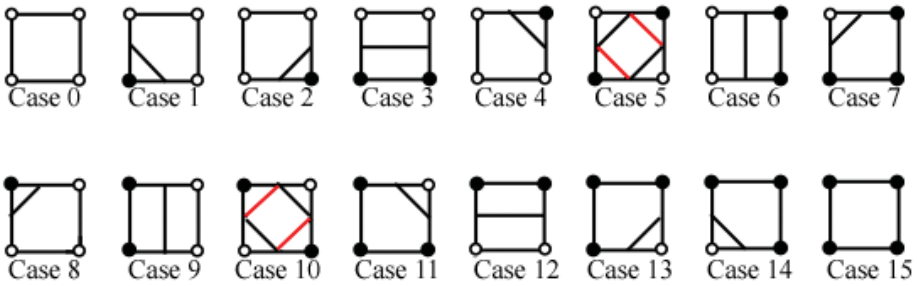


Рисунок 3.1. Набор структур единичных фрагментов для поиска контура.

Пример построения «карты поверхности» приведен на рисунке 3.2. Картой поверхности является основание трехмерной проекции интенсивностей на изображении. Полученные признаки удельного количества и площади аномалий рассчитываются как соответствующая средняя величина с нормировкой на площадь фотографии. Затем значения для каждого объекта наряду со средней яркостью изображения заносятся в базу данных «объекты-признаки».

Гиперпараметрами работы оптического регистратора являются: окно сглаживания для фильтра [1, 30] и «константа контуров» [0,1 – 2,0] - параметр работы алгоритма «совпадающих квадратов», определяющий относительную площадь единичного фрагмента на изображении (рисунок 3.3).

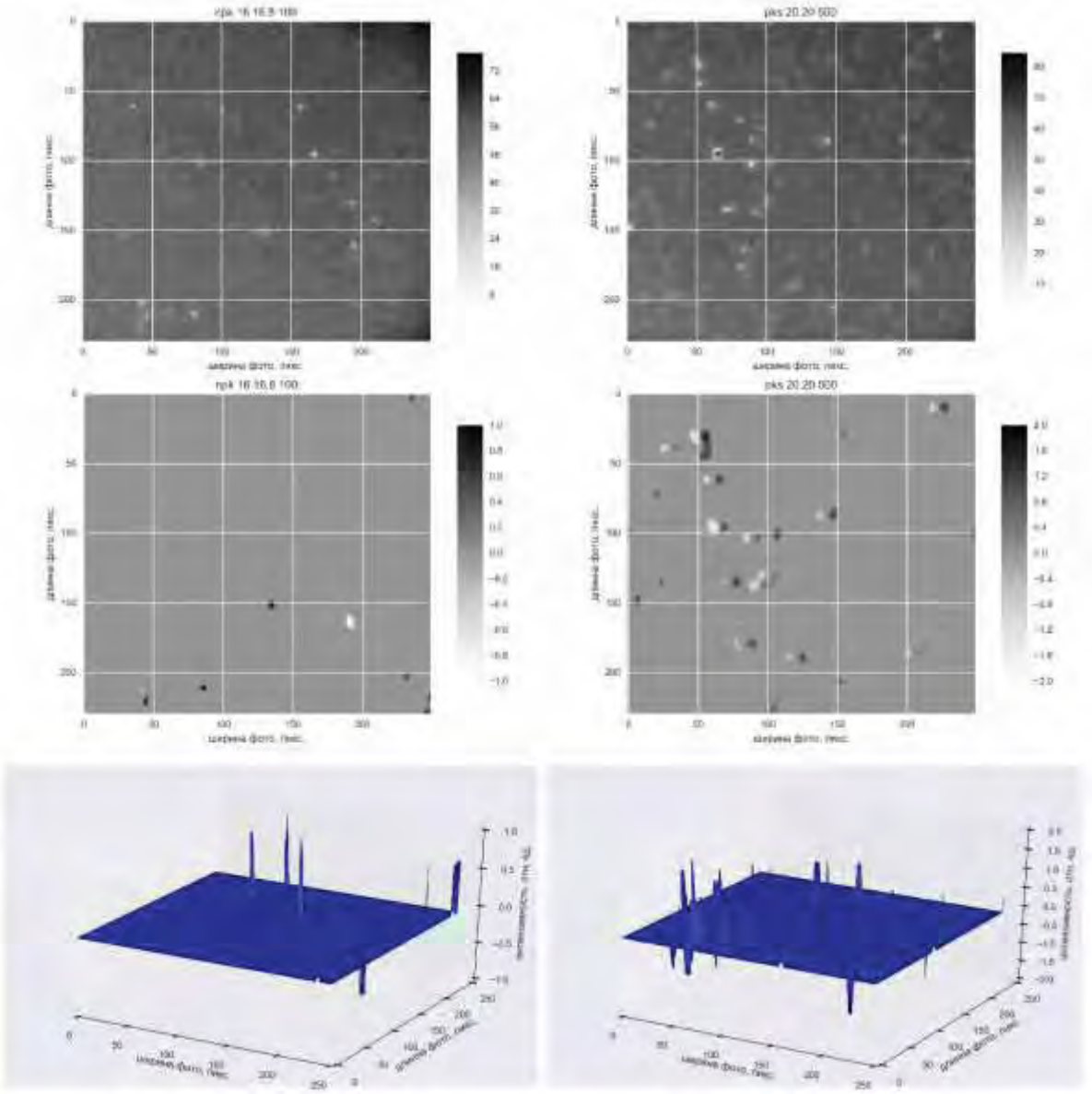


Рисунок 3.2. Построение карты поверхности для различных объектов, рассмотренных в работе: исходное изображение в градация серого (яркость), продифференцированное и сглаженное изображения (карта поверхности) и трехмерное изображение карты поверхности.

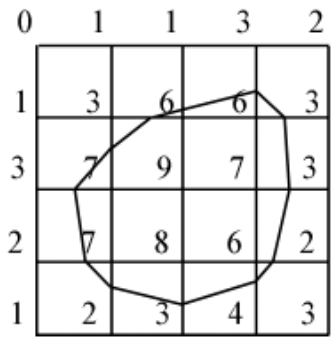


Рисунок 3.3. Контур на изображении, разбитый на единичные фрагменты.

3.2.2 Выделение «физико-химических» признаков пробы из РФ-спектра

Предложен алгоритм и стадии оптимизации для расчета «физико-химических» признаков по РФ-спектру исследуемого объекта. Стадии алгоритма приведены в таблице 3.1.

Таблица 3.1. Выделение признаков из РФ спектра

№ стадии	Описание стадии	Основные уравнения
1	Измерение спектра	-
2	Сглаживание спектра для нивелирования аппаратных шумов	Среднее: $y_i = \frac{1}{2w}\sum_{j=i-w}^{i+w} y_j$ где: *w* – полуширина окна [1, 30]*, *y* – сигнал
		Медиана: $y_i = median_{j=i-w}^{i+w}(y_j)$ где: *w* – полуширина окна [1, 30]*, *y* – сигнал
		Алгоритм Савицкого-Голея $y_i{}^l = \sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} 1 + P^{or}[C_i y_{j+i}]$ $\frac{m+1}{2} \le j \le n - \frac{m-1}{2}$ где: $y_i{}^l$ – исправленное значение спектра, y_i – исходное значение спектра, P^{or} – полином, *or* – заданная степень полинома [2, 5]*, C_i – оптимизационный коэффициенты, *n* – количество точек в окне, *m* – полуширина окна [7, 31]*.
		Алгоритм дискретного Фурье преобразования с отсечением высоких частот $Y(u) = \frac{1}{n}\sum_{i=0}^{n-1} y_i\, e^{\frac{-j2\pi ui}{n}}$ где: *Y(u)* – преобразованный спектр состоящий из амплитуд и фаз, $j = \sqrt{-1}$ и *u = 0, ..., n-1*, y_i – дискретный спектр, *n* – количество каналов, $Y^*(u) = Y(u)H(u)$ *H(u)* – фильтрующая функция $H(u) = \begin{cases} 1, & u \le u_{th} \\ 0, & u > u_{th} \end{cases}$ u_{th} – порог отбрасываемых частот [0.01, 2.00] Гц*. После фильтрации частот: $y_i = \sum_{i=0}^{n-1} Y(u)\, e^{\frac{j2\pi ui}{n}}$ подбирая параметры отрезаемых частот можно добиться фильтрации аппаратных шумов.

3	Выделение фоновой линии и определение положения и ширины характеристических линий	Алгоритм нулевого фильтра: $$y_i{}^l = \sum_{k=-v-\frac{w}{2}}^{v+\frac{w}{2}} h(k) y_{i+k}$$ $$h(k) = \begin{cases} -\frac{1}{2v}, & -v-\frac{w}{2} \le k < -\frac{w}{2} \\ \frac{1}{w}, & -\frac{w}{2} \le k \le \frac{w}{2} \\ -\frac{1}{2v}, & \frac{w}{2} < k \le \frac{w}{2}+v \end{cases}$$ где: $y_i{}^l$ – исправленное значение спектра, y_i – исходное значение спектра, $h(k)$ – коэффициент, v – первый параметр окна [2, 20]*, w – второй параметр окна , характеризующий «сдвиг» относительно первого (должен быть четным) [2, 20]*.
		$\mathrm{B} = \mathrm{S} - \mathrm{B}_0$ где: B – вектор каналов базовой линии, S – вектор каналов спектра, B_0 – вектор каналов пиков. Вычитаемые значения заменяются краевыми.
		По маске найденных пиков из спектра выделяется фоновая составляющая
5	Аппроксимация базовой линии	Расчет площади и максимума интенсивности аппроксимированной фоновой составляющей
6	Выделение характеристических линий	Дифференциальный. Исходный спектр дифференцируется и итеративно сглаживается оптимальным фильтром - [0, 10]* итераций. Далее определяются все локальные максимумы, превышающие заданный порог [5, 25]*.
		Обратный алгоритм «нулевого фильтра». По полученной маске пиков проводиться несколько итераций сверх сглаживания наиболее оптимальным типом алгоритма – [0, 10]* итераций и выделяются все локальные максимумы большие 0.
7	Расчет аналитического сигнала найденных линий	Аппроксимация линии по Гауссу: $$I_g = Ae^{-\frac{(I-\mu)^2}{2\sigma^2}}$$ где: I_g – интенсивность гауссианы, I – интенсивность линии спектра, A - высота, μ - середина и σ – ширина линии спектра и последующий расчет интенсивности и площади гауссианы.
		Вычисление интенсивности линии как среднего из трех максимальных компонент.

* подбираемый параметр с указанием диапазона оптимизации

В качестве признаков используются определенные в результате работы алгоритмов интенсивности характеристических линий, максимальная интенсивность фоновой линии и площадь под фоновой линией. Предложенный

набор алгоритмов полностью автоматизирован и по совокупности стадий является уникальным.

3.3 Создание единой базы данных «объекты-признаки»

После выделения основных физических и химических признаков пробы составляется матрица «объекты-признаки», являющаяся основной для дальнейшей статистической обработки и АБД. Данная матрица должна включать в себя следующий набор признаков (таблица 3.2).

Таблица 3.2. Учитываемые признаки объектов

Тип признака	**Тип принимаемых значения**	**Источник получения**	**Количество признаков данного типа**
Интенсивности химических элементов в спектре	вещественный	РФ спектр	порядка 10
Содержание основных питательных (N, P, K) и прочих элементов (S, Zn)	вещественный	паспорт*	3 - 10
максимальная фракция	категориальный	оптический	1
предварительная сушка	бинарный	паспорт*	1
обработка кондиционирующей добавкой	вещественный или бинарный (факт наличия добавки)	оптический или паспорт	1
удельное количество артефактов карты поверхности	вещественный	оптический	1
удельная площадь артефактов карты поверхности	вещественный	оптический	1
тип и марка пробы	категориальный	паспорт*	1

* получаем по известным данным о пробе или с использованием алгоритмов классификации/регрессии по известным данным

Далее проводиться предобработка признаков, которая при необходимости включает в себя:

1. Предобработку категориальных и бинарных признаков – «бинарное кодирование». Рассмотрим общий случай. Пусть задано $x_j = (c_1, \dots, c_n)$ – n значений категориального признака x_j. Для кодировки вводят n новых

бинарных признаков: $b_1(x), \dots, b_n(x)$. Таким образом значение бинарного признака равно 1, когда на данном объекте значение категориального признака $f_j(x)$ равно c_i:

$$b_i(x) = [f_j(x) = c_i]$$

В результаты один категориальный признак заменяется n бинарными.

2. Балансировку классов исследуемых объектов.
   a. «Undersampling» – отбрасываем часть объектов из больших классов.
   b. «Oversampling» – добавляем объекты в меньшие классы (дублируем или задаем погрешность).

   Количество удаляемых объектов является гиперпараметром.

3. Масштабирование вещественных признаков:
   a. стандартизация (без сохранения информации о дисперсии):

$$x_i^j = \frac{x_i^j - \mu_j}{\sigma_j}$$

$$\mu_j = \frac{1}{l}\sum_{i=1}^{l} x_i^j$$

$$\sigma_j = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(x_i^j - \mu_j)^2}$$

   где: $x^j{}_i$ – значение признака, μ_j – среднее по выборке, σ_j – корень стандартного отклонения данных по выборке.

   b. масштабирование на размах (отрезок [0, 1], с сохранением информации о дисперсии):

$$x_i^j = \frac{x_i^j - m_j}{M_j - m_j}$$

$$m_j = \min(x_1^j, \dots, x_i^j)$$

$$M_j = \max(x_1^j, \dots, x_i^j)$$

   где: $x^j{}_i$ – значение признака, m_j –минимум по выборке, M_j – максимум по выборке.

По результатам работы данного блока получаем матрицу «объекты-признаки» где все параметры подготовлены для дальнейшей классификации и/или регрессии.

3.4 Построение моделей регрессии.

Для предсказания и поиска структуры в вещественных непрерывных признаках используются алгоритмы регрессии, которые в общем случае делятся на линейный и нелинейные. Учитывая относительную линейность зависимость интенсивностей характеристических линий от концентраций и возможность использования нелинейных методов понижения размерности акцент будет сделан на линейных методах регрессии. К тому же данные методы наиболее распространены в аналитической практике.

Рассматриваемые алгоритмы в задачах регрессии имеют общий вид:

$$a(x) = \sum_{j=1}^{d+1} w_j x^j = \langle w, x \rangle$$

где: w_j - веса, x_j – признаки, где в каждый объект добавлен свободный член (признак равный 1), $\langle w, x \rangle$ – скалярное произведение весов модели и признаков (ответ модели).

В качестве меры ошибки рассмотрим наиболее распространенную и универсальную средне квадратичную ошибку (СКО). Функционал ошибки имеет вид:

$$Q(a, x) = \frac{1}{l} \sum_{i=1}^{l} (\langle w, x \rangle - y_i)^2$$

Тогда задача поиска оптимально регрессионной модели в матричном виде:

$$Q(w, X) = \frac{1}{l} \|Xw - y\|^2 \longrightarrow \min_w$$

Данная задача имеет так же и аналитическое решение:

$$w = (X^T X)^{-1} X^T y$$

где: X – матрица признаков, y – вектор ответов модели, w – вектор весов модели, l – количество итераций.

В современной химии, как правило, используется именно аналитический вариант решения регрессионной задачи. Однако данный подход обладает следующими недостатками:

- для нахождения решения требуется вычислять обратную матрицу. Данная операция требует порядка d^3 операций для d признаков и является вычислительно сложной.
- Данный метод не применим когда матрица плохо обусловлена ($cond(A) = \|A\| \cdot \|A^{-1}\| > 10^3$, например, когда признаки являются линейно зависимыми).

По этому, в представленной работе применен оптимизационный подход к решению, основанный на множестве итераций по уменьшению функционала ошибки, основанный на методе градиентного спуска:

$$w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1}, X)$$

где: w_t – значение весов на t-итерации, w_{t-1} – предыдущее приближение ($w_0 = 0$), η_t – шаг градиентного спуска, $\nabla Q(w^{t-1}, X)$ – направление градиента.

Итерации завершают, когда наступает сходимость:

$$\|w^t - w^{t-1}\| < \varepsilon$$

В многомерном случае:

$$\nabla_w Q(w, X) = \frac{2}{l} X^T (Xw - y)$$

Существует варианты расширения моделей регрессии. Например наиболее распространенный подход основан на использовании спрямляющих пространств:

1. добавление квадратичных признаков:

$$(x_1, \dots, x_d) \to (x_1, \dots, x_d, x_1^2, \dots, x_d^2, x_1 x_2, \dots, x_{d-1} x_d),$$

при этом увеличивается риск переобучения модели.

2. добавление полиномиальных признаков: аналогично предыдущему, но для более высоких степеней. Еще сильнее увеличиваем риск переобучения.
3. логарифмирование (и любые другие нелинейные функции)

$$x_i \to \ln(x_i + 1),$$

$$x_i \to \ln(|x_i| + 1)$$

Однако, учитывая структуру данных, в нашей работе используются в основном линейные методы регрессии по множеству компонентов (признаков) объекта.

3.5 Построение моделей классификации.

Для бинарных, категориальных и дискретных вещественных признаков в работе с большими данными широко распространены различные алгоритмы классификации, начиная от простейший линейных разделяющих поверхностей и «решающих пней» (алгоритмов да/нет), заканчивая сложными нейронными сетями. Учитывая достаточно простую природу наших данных (ожидаем зависимости, близкие к линейным) и отсутствия в них сложных скрытых структур алгоритмы нейронных сетей в работе рассматриваться не будут. Акцент сделан на следующих классификаторах:

1. Линейные методы. Решаем задачу линейной регрессии, описанную ранее, с поиском функции, задающей разделяющую поверхность и отображающей объекты классификации в вещественное пространство:

$$g\big(\mathbb{E}(y|x)\big) \approx \langle w, x\rangle$$

$$g\colon (0,1) \to \mathbb{R}$$

$$a(x) = sign \sum_{j=1}^{d+1} w_j x^j = sign\langle w, x\rangle$$

2. Алгоритмы случайного леса – отражают нелинейную логику принятия решений. Основная идея состоит в обучении композиции алгоритмов решающих деревьев (да/нет алгоритм определенной глубины) и усреднении полученных ответов:

$$a(x) = sign\frac{1}{N}\sum_{n=1}^{N} b_n(x), \text{ для N алгоритмов}$$

Рассмотрение более подробного математического аппарата случайного леса выходит за рамки данной работы.

3. Наивный Байесовский классификатор. Пусть некоторый объект имеет вектор признаков x. Тогда класс объекта определяется как максимизация вероятности принадлежности объекта к группе класса:

$$a(x) = argmax_y P(y|x) = argmax_y P(x|y)P(y)$$

Более подробное рассмотрение Байесовских методов так же выходит за рамки представленной работы.

В качестве метрик качества алгоритмов классификации выступает доля неправильных ответов, минимизируемая сверху:

$$Q(a,x) = \frac{1}{l}\sum_{i=1}^{l}(a(x) \neq y_i)$$

$$Q(a,x) = \frac{1}{l}\sum_{i=1}^{l}(y_i\langle w, x_i\rangle < 0)$$

$$Q(a,x) < \tilde{Q}(a,x) = \frac{1}{l}\sum_{i=1}^{l}\tilde{L} \to \min_a$$

$$\tilde{L} = \log_2(\exp(-y_i\langle w, x_i\rangle))$$

Полученная оценка является гладкой и появляется возможность использовать итерационные методы оптимизации, описанные в предыдущем пункте работы, такие, как градиентный спуск.

3.6 Способы оценки качества работы и оптимизации алгоритмов

Для каждого рассмотренного алгоритма актуальна проблема так называемого «переобучения» - построения слишком сложных моделей, хорошо описывающих тестовые данные, но совершенно не применимые для реальных. Особенно актуальна данная проблема для методов множественной регрессии. В данном случае, для устранения описанной проблемы, в автоматическом режиме используются методы регуляризации для отбора наиболее значимых признаков и методы кросс-валидации для усредненной оценки качества работы алгоритмов.

Стратегии кросс-валидации применяются для всех типов алгоритмов, рассмотренных в данной работе.

3.6.1 Регуляризация

Обозначенная ранее проблема «переобучения» алгоритмов характеризуется большими весами при признаках модели и часто возникает при наличии линейно зависимых параметров (мультиколлинеарность). Для разрешения данной проблемы используют минимизацию не функционала ошибки, а функционала ошибки с добавлением поправочного коэффициента – регуляризатора. В работе используются два наиболее распространенных способа регуляризации:

1. Абсолютный регуляризатор (Lasso, L1). Задача оптимизации включает в себя оптимизацию СКО и сумму модулей всех коэффициентов, умноженную на коэффициент регуляризации (λ):

$$Q(w,X) + \lambda\|w\| = Q(w,X) + \lambda\sum_{j=1}^{d}|w_j| \to \min_w$$

$$w_* = argmin_w(\frac{1}{l}\sum_{i=1}^{l}(\langle w, x_i\rangle - y_i)^2 + \lambda\sum_{j=1}^{d}|w_j|)$$

   При этом некоторые веса обнуляются и таким образом происходит автоматический отбор значимых признаков модели.

2. Квадратичный регуляризатор (Ridge, L2). Задача оптимизации включает в себя оптимизацию СКО и сумму квадратов всех коэффициентов:

$$Q(w,X) + \lambda\|w\|^2 = Q(w,X) + \lambda\sum_{j=1}^{d}{w_j}^2 \to \min_w$$

$$w_* = argmin_w(\frac{1}{l}\sum_{i=1}^{l}(\langle w, x_i\rangle - y_i)^2 + \lambda\sum_{j=1}^{d}w_j^2)$$

где: λ – коэффициент регуляризации, чем больше – тем ниже сложность модели.

При этом модель L2 регуляризации имеет так же и аналитическое решение:

$$w_* = (X^T X + \lambda I)^{-1} X^T y$$

С использованием регуляризации вводиться штраф за слишком большие веса модели, а коэффициент λ является гиперпараметром и подлежит оптимизации в диапазоне [0, 1].

3.6.2 Кросс-валидация

Для независимой и устойчивой оценки качества работы алгоритмов используется понятие отложенной выборки и связанное с ним понятие кросс-валидации. Отложенная выборка представляет из себя часть объектов, не участвующих в построении модели и использующаяся только для вычисления метрик качества работы алгоритма. Однако, для более устойчивого значения, используют среднее по оценке на множестве отложенных выборок, которые формируются различными путями – так называемые стратегии кросс-валидации. Рассмотрим основные параметры описанных подходов оценки метрик качества.

1. Отложенная выборка. Исходная выборка разбивается на две части – для обучения и тестирования алгоритма. Как правило используются разбиение в соотношениях: 70/30 или 80/20. Преимуществом данного подхода является скорость (обучать алгоритмы нужно один раз), однако результат не устойчив и сильно зависит от качества разбиения.
2. Кросс-валидация. Более системный подход. Выборка делиться на k блоков примерно одинакового размера. Далее каждый из блоков используется в качестве тестового а все остальные – в качестве обучающих. По результатам показатель качества усредняют. Обычно k = 3, 5 и 10. Существует несколько стратегий разбиения данных по кросс-валидации:
    a. Простое разбиение на *k* подвыборок (одинарная кросс-валидация);
    b. Одинарная кросс-валидация с перемешиванием объектов внутри выборки;
    c. Одинарная кросс-валидация с сохранением распределения классов (стратификация);

В работе используется последний вариант с сохранением распределения классов и дополнительном перемешивании данных по всей выборке объектов. Данная стратегия на наш взгляд является универсальной и позволит добиться устойчивых оценок работы алгоритмов. Таким образом, в данной работе метрики качества оценивались по кросс-валидации как среднее по десяти наборам тестовых данных: случайно отобранные 30 % объектов из общей выборки с сохранением распределения целевых классов.

3.6.3 Задача оптимизации

Оптимизация рассмотренных алгоритмов заключается в подборе различных гиперпараметров. При подборе оптимальных условий полученный алгоритм наиболее хорошо описывает исследуемые данные и, как следствие, лучше всего минимизирует функционалы ошибки или обладает наилучшим значением метрики качества. Данный подход заключается в программном переборе параметров по «сетке» - программной структуре, включающей в себя параметры для оптимизации и шаг их изменения. Примерами гиперпараметров, оптимизируемых в данной работе являются:

- параметр регуляризации λ (при использовании регуляризаторов);
- количество итераций обучения;
- параметры алгоритмов предобработки спектров (таблица 3.1);
- количество вершин в алгоритмах случайного леса и т.д.

Все оптимизируемые гиперпараметры для каждого типа алгоритма приведены в таблице 3.3.

Таблица 3.3. Основные оптимизируемые гиперпараметры, в скобках приведены обозначения из Python 2.7

Алгоритм	**Гиперпараметр**	**Значение гиперпараметра**
Линейная классификация с градиентным спуском (SGDClsiifier)	Функция потерь (loss)	hinge – метод опорных векторов (классификация с максимальным зазором), log – логистическая (вероятностная), squared_loss – СКО, modified_huber – функция потерь Хьюберта, более устойчива к выбросам в данных чем СКО.
	Количество итераций градиентного спуска (n_iter)	1000 – 15000
Линейная классификация с градиентным спуском (SGDClsiifier) с L1 или L2 регуляризацией	Функция потерь (loss)	hinge, log, squared_loss, modified_huber
	Количество итераций градиентного спуска (n_iter)	1000 – 15000
	Коэффициент регуляризации (alpha)	0,0001 - 1
Случайного леса (RandomForestClassifier)	Количество решающих деревьев (n_estimators)	2 – 100
	Ограничение максимального количества признаков (max_features)	Нет (None), квадратичное (sqrt), логарифмическое (log2)
	Ограничения глубины решающего дерева (max_depth)	Нет (None), 2 – 13
	Использование бутстрапа (bootstrap)	Да (True), нет (False)
	Балансировка классов (class_weight)	Да (balanced), нет (None)
Наивный байесовский классификатор (MultinomialNB)	-	-
Линейная регрессия по множеству компонент	-	-
Линейная регрессия по множеству компонент с L1 или L2 регуляризацией	Нормализация данных (normalize)	Да (True), нет (False)
	Количество итераций градиентного спуска (n_iter)	1000 – 15000
	Параметр регуляризации (alpha)	0 – 1

Сравнение качества работы алгоритмов происходит с использованием выбранной ранее стратегии кросс-валидации, по которой рассчитываются усредненные значения различных метрик качества.

3.6.4 Метрики качества

Метрики качества это приближения для различных функционалов ошибки алгоритмов, которые несут в себе определенный физический смысл. Так, различные метрики могут использоваться для:

- задания функционала ошибки при обучении алгоритма;
- подбора гиперпараметров;
- оценки итоговой модели: пригодна ли модель для решения задачи.

Далее приведены наиболее распространенные и универсальные метрики качества для алгоритмов классификации и регрессии. Прочие метрики вводятся по ходу работы.

3.6.4.1 Задачи регрессии

Среднеквадратичная ошибка:

$$MSE(a.X) = \frac{1}{l}\sum_{i=1}^{l}(a(x) - y_i)^2$$

Описывает насколько сильно предсказанные значения отклоняются от истинных. Такой параметр легко оптимизировать, однако данный функционал сильно штрафует за большие ошибки, что приводит к настройке алгоритма на выбросы.

Средняя абсолютная ошибка:

$$MAE(a.X) = \frac{1}{l}\sum_{i=1}^{l}|a(x) - y_i|$$

Описывает абсолютное отклонение предсказанных данных от истинных. Данный функционал сложнее минимизировать, поскольку производная не определена в 0, однако он более устойчив к выбросам.

Коэффициент детерминации:

$$R^2(a.X) = 1 - \frac{\sum_{i=1}^{l}(a(x)-y_i)^2}{\sum_{i=1}^{l}|y_i-\bar{y}|}, \bar{y} = \frac{1}{l}\sum_{i=1}^{l} y_i$$

$$0 \leq R^2 < 1$$

Показывает, какую долю дисперсии во всем целевом векторе ответов модель смогла объяснить. Хорошо работает для линейных моделей и является квадратом коэффициента корреляции Пирсона, который используется для описания силы линейной взаимосвязи между двумя переменными.

3.6.4.2 Задачи классификации

Для задач классификации вводится понятие доли правильных ответов алгоритма для определенного класса:

$$accuracy(a, X) = \frac{1}{l}\sum_{i=1}^{l}[a(x) = y_i]$$

Как правило используют усредненное значение по всем классам. Наиболее очевидная метрика, однако обладающая рядом недостатков:

- завышение при несбалансированных выборках;
- не учитывает разные цены ошибки.

Для устранения описанных проблем вводят понятие «матрицы ошибок»:

	y = 1	**y = -1**
a(x) = 1	верное положительное (ВП)	не верное положительное (НП)
a(x) = -1	не верное отрицательное (НО)	верное отрицательное (ВО)

По описанной матрице рассчитывают точность и полноту работы алгоритма:

$$precision(a, X) = \frac{\text{ВП}}{\text{ВП} + \text{НП}}$$

$$recall(a, X) = \frac{\text{ВП}}{\text{ВП} + \text{НО}}$$

Более универсальной метрикой (но менее очевидной) является гармоническое среднее или F-мера (F1):

$$F = \frac{2 * precision * recall}{precision + recall}$$

Данная метрика обобщает точность и полноту, является гладкой и легко оптимизируется.

3.7 Кластеризация, понижение размерности и представление многомерных данных

Используя описанный математический аппарат становиться возможным полное физико-химическое описание известных объектов с использованием заранее созданной матрицы «объекты-признаки». Однако, зачастую, в производстве требуется проанализировать ранее не встречавшийся объект или группу объектов. Оценить, насколько они отличаются от известных объектов, сколько имеют аномалий и как связаны друг с другом, помогают алгоритмы кластеризации и визуализации на плоскости. Для этого в работе используются метрические алгоритмы и методы понижения размерности.

Как правило, для данных методов вводиться новое понятие метрики качества, как функции, задающей расстояние в метрическом пространстве:

- Евклидова метрика: $p(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$;
- Манхэттенская метрика: $p(x, y) = \sum_{i=1}^{n}|x_i - y_i|$;
- метрика Минковского (обобщение Евклидовой и Манхэттенской метрики):

$$p(x, y) = \left(\sum_{i=1}^{n}|x_i - y_i|^q\right)^{\frac{1}{q}}$$

Описанный параметр позволяет не только рассчитать оптимальный путь от одного класса до другого, но так же и получить количественную оценку величины класса и степени близости различных групп объектов. Для этого могут использоваться:

- коэффициент корреляции Пирсона: $\frac{\sum_{i=1}^{n}((x_i-\bar{x})(y_i-\bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$
- скалярное произведение: $\sum x_i y_i$
- коэффициент Дайса: $\frac{2\sum x_i y_i}{\sum {x_i}^2+\sum {y_i}^2}$
- косинусная мера: $\frac{\sum x_i y_i}{\sum {x_i}^2 \sum {y_i}^2}$
- коэффициент Жаккара: $\frac{\sum x_i y_i}{\sum {x_i}^2+\sum {y_i}^2-\sum x_i y_i}$

В работе в основном используется Евклидова метрика и коэффициент корреляции Пирсона, если не оговорено обратное, поскольку зависимости в данных предположительно близки к линейным.

3.7.1 Кластеризация

Задача кластеризации ставиться как восстановления отображения «объекты – признаки»: $x \mapsto y$. И представляет собой итеративный процесс, по сути своей мало отличающийся от классификации. Метрики качества, как правило являются производными от среднего расстояния между объектами в кластере (F_0) и среднего расстояния между объектами разных кластеров (F_1):

$$F_0 = \frac{\sum_{i<j}[y_i = y_j]p(x_i, x_j)}{\sum_{i<j}[y_i = y_j]} \to min$$

$$F_1 = \frac{\sum_{i<j}[y_i \neq y_j]p(x_i, x_j)}{\sum_{i<j}[y_i \neq y_j]} \to max$$

$$\frac{F_0}{F_1} \to min$$

Существует целый ряд алгоритмов кластеризации, наиболее распространенные из которых приведены далее.

3.7.1.1 Взвешенный метод ближайших соседей (k-средних, k-means, kNN)

Объект относятся к том классу, к которому принадлежит большинство из его k ближайших соседей с определенными весами, зависящими от расстояния до соседа:

$$a(x) = argmax_{y \in Y} \sum_{i=1}^{k} [x_i = y] w(x_i),$$

- $w(d) = \frac{1}{(d+a)^b}$
- $w(d) = q^d$, 0 < q < 1

где: $w(x_i)$ – функция расчета расстояния до «соседа».

В качестве параметров метода выступает количество кластеров. Метод хорошо масштабируем на различные величины выборок, однако неустойчив по отношению в несбалансированным классам. Применим для выпуклых кластеров примерно одинакового размера. В качестве метрики расстояния использует Евклидову метрику.

3.7.1.2 EM-алгоритм

Данный метод заключается в максимизации правдоподобия вероятности принадлежности точки к кластеру. Основан на разделении смеси распределений. Заключается в последовательном выполнении двух шагов:

1. E-шаг: вычисляются вспомогательные переменные.

$$g_{ji} = p(\theta_i | x_i) = \frac{w_j p_j(x_i)}{p(x_i)}$$

Данные параметры фиксируются для упрощения задачи максимизации.

2. M-шаг: при зафиксированных параметрах g_{ji} решение задачи максимизации правдоподобия может быть найдено согласно:

$$w_j = \frac{1}{N} \sum_{i=1}^{N} g_{ji} \qquad \theta_j = argmax_\theta \sum_{i=1}^{N} g_{ji} \ln \varphi(\theta, x)$$

Объект относиться к кластеру j для которого максимально значение p(Qj|x).

Например для смеси гауссовых распределений:

$$p(x) = \frac{1}{2} N\left(\binom{4}{0}, 1\right) + \frac{1}{2} N\left(\binom{8}{0}, 1\right)$$

$$N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

1. E-шаг:

$$g_{ji} = p(\theta_i|x_i) = \frac{w_j p_j(x_i)}{p(x_i)}$$

2. M-шаг:

$$w_j = \frac{1}{N}\sum_{i=1}^{N} g_{ji} \qquad \mu_j = \frac{1}{Nw_j}\sum_{i=1}^{N} g_{ji}x_i \qquad \sigma_j = \frac{1}{Nw_j - 1}\sum_{i=1}^{N} g_{ji}(x_i - \mu_j)^2$$

Параметром метода является количество компонент. Метод хорошо работает на больших массивах данных и обладает средней масштабируемостью. В качестве метрики используется обобщенная Евклидовая метрика.

3.7.1.3 Методы, основанные на плотности точек

В основе данной группы методов лежит рассмотрение окрестности каждой точки по которой она классифицируется как:

- основная – в окрестности много других точек;
- пограничная – в окрестности мало других точек;
- шумовая – в иных случаях.

Параметрами метода являются радиус окрестности и количество соседей. Данный метод обладает хорошей масштабируемостью и может отсеивать выбросы. В качестве метрики используется Евклидова метрика.

Исходя из предполагаемой линейности данных и того факта, что заранее известно количество кластеров в работе в основном используется алгоритм k-средних, если не оговорено обратное.

3.7.2 Методы понижения размерности

Следующей задачей, решаемой АБД является понижение размерности признаков объектов. Данный класс алгоритмов используется для отсева шумовых признаков, ускорения обучения моделей и качественной визуализации. Учитывая, что количество признаков одного объекта в данной работе не превышает 100, понижение размерности в основном будет использоваться для визуализации и

кластеризации объектов. Методы понижения размерности в общем случае делятся на два типа.

3.7.2.1 Одномерный отбор признаков

Основаны на оценки связи каждого признака с целевой переменной. Наиболее просты и наглядны. Для этого существует несколько подходов:

1. Оценка корреляции:

$$R_j = \frac{\sum_{i=1}^{l}(x_{ij} - \bar{x_j})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{l}(x_{ij} - \bar{x_j})^2 \sum_{i=1}^{l}(y_i - \bar{y})^2}}$$

Чем больше по модулю корреляция, тем больше информативность данного признака и линейность его связи с целевой переменной.

2. Использование бинарной классификации по всем признакам. Для этого обычно рассматривают взаимосвязь всех признаков друг с другом на плоскости и оценивают предсказательную способность для определения целевой переменной.

Данные группы методов используются при построении карт линейных корреляций и бинарных классификаций. Однако рассмотренные методы не всегда информативны, поэтому зачастую используют отбор признаков на основе каких-либо моделей регрессии или классификации.

3.7.2.2 Понижение размерности данных на основе модели

Как упоминалось ранее данная группа алгоритмов используется в обучении различных классификаторов или регрессии для проецирования всей совокупности свойств в пространство заданной размерности. Например, так обеспечивается качественная визуализация множества компонент на плоскости. Исходя из предполагаемой линейности данных в работе будут рассмотрены следующие подходы.

1. Метод случайных проекций.

Является линейным подходом, в котором каждый новый признак представляет собой линейную комбинацию исходных:

$$z_{ij} = \sum_{k=1}^{D} w_{jk} x_{ik}$$

где: z_{ij} – значение нового признака j на объекте выборки i, D – количество исходных признаков, x_{ij} – значение исходного признака j на объекте выборки i, d – число новых признаков.

При этом веса признаков генерируются случайно, например из нормального распределения:

$$w_{ik} \sim N(0, \frac{1}{d})$$

Описанный подход прост в реализации, однако неустойчив и обеспечивает качественное представление только сильно отличающихся данных.

2. Метод главных компонент.

Так же является линейным алгоритмом. Задача данного метода – матричное разложение (необходимо представить исходную матрицу признаков в виде произведения матрицы новых признаков и ортогональной матрицы весов, которые будут иметь меньший ранг). Обсуждение математического аппарата данного метода выходит за рамки представленной работы, программная реализация данного алгоритма представлена в приложении А.

3. Нелинейные методы понижения размерности.

Примерами таких методов являются метод многомерного шкалирования и TSNE (t-распределенных стохастических соседних вложений). Однако данные методы сложны для интерпретации полученных результатов, а обсуждение их математического аппарата выходит за рамки данной работы.

3.8 Заключение

В результате обзора основных методов анализа больших данных подобраны математические алгоритмы для решения поставленных в работе задач. Так, после выделения и предобработки признаков из РФ-спектров и оптических изображений проб, предлагается оценить информативность выделенных свойств по их способности к:

- классификации физических свойств пробы (фракции, типа и марки);
- регрессии по концентрации основных питательных элементов (N, P, K) и некоторых других (S, Zn);
- общему представлению структуры данных (кластеризация и визуализация).

В результате исследования планируется составить единую матрицу «объекты-признаки», содержащую в себе все проанализированные пробы и все выделенные, согласно описанным алгоритмам, физические и химические свойства. По полученной базе данных становиться возможным применение различных алгоритмов поиска неизвестных химических и физических параметров, характеризующих качество выпускаемой продукции, в том числе и по не явным признакам. Так, например, предполагается оценить:

- марку удобрений по азоту;
- гранулометрический состав и фактор формы гранул;
- качество обработки кондиционирующими добавками;
- расчет проводимости разбавленных растворов удобрений (так называемый «солевой индекс» [116]).

Дополнительно, для каждой математической стадии, будет проведена процедура оптимизации и рассчитаны соответствующие метрики качества.

4 Постановка экспериментальной части

4.1 Использованная аппаратура и реактивы

В представленной работе использованы:

- Отечественные ЭД РФ-спектрометры РЕАН и X-Spec, обладающие рентгеновскими трубками прострельного типа мощностью 50 Вт и 10 Вт соответственно, с Mo и Th анодами соответственно, производства АО «Научные приборы».
- Отечественный сканирующий рентгеновский микроскоп Рам-30μ мощностью 50 Вт с Rh трубкой прострельного типа, производства АО «Научные приборы».
- Цифровая USB видеокамера с характеристиками:

Параметр	Значение
Разрешение	не менее 640x480
Фокусное расстояние	2,8 – 12 мм
Датчик	½. 7`` CMOS

- Весы аналитические 1 класса точности.
- Весы лабораторные общего назначения.
- Пресс лабораторный ПЛГ-12 с усилием не менее 12 т.
- Пресс-формы диаметром 20 мм с фигурными пуансонами в комплекте.
- Размольное устройство любого типа, обеспечивающее требуемый дисперсный состав пробы после измельчения (не более 0,500 мм и 0,100 мм).
- Набор плетеных сит с сеткой № 0,1 и 0,5.
- Шкаф сушильный электрический лабораторный с точностью автоматического регулирования температуры не более ± 2 ^{0}C в диапазоне температур (80 – 250) ^{0}C.
- Влагомер типа Sartorius MA 150 обеспечивающий работу в диапазоне температур (80 – 250) ^{0}C.
- Кондуктометр с возможностью измерения проводимости до 15 мС/см.

- Бумажный фильтр белая лента.
- Делитель проб лабораторный.
- Роторная ножевая мельница типа Ika Basik обеспечивающая размол до 100 мкм.

Реактивы:

- кислота борная по ГОСТ 18704-78 или ГОСТ 9656-75;
- спирт этиловый по ГОСТ 18300-87 или ГОСТ Р 51562-2000.
- дистиллированная вода.

Все используемые реактивы должны соответствовать квалификации ч.д.а. Допускается использование реактивов более высокой квалификации или импортных аналогов с квалификацией не ниже ч.д.а.

В работе использовали набор рабочих проб с предприятий АО «ФосАгро». Их составы приведены ниже (таблица 4.1).

Таблица 4.1. Типы реальных проб, использованных в работе

№	Тип объекта	Марка	Количество
1	NPK(S)	4-30-15(16)	30
2	NPK(S)	0-20-20(5)	30
3	NP(S)	12-40(10)	30
4	NPK	15-15-15	40
5	NPK	16-16-8	40
6	NP(S)+S+Zn	12-40(6)+3+1	70
7	МАФ (NP)	12-52	30

4.2 Процедуры измерений и пробоподготовки

Общие условия измерения рентгеновских спектров для различных объектов составили 25 кВ, 100 мкА и 50 с, при которых мертвое время детектора для разных объектов не превысило 20 %. Для прочих частных случаев условия приведены по тексту работы. Обоснование использования данных параметров приведено в главе 5 настоящей работы.

Согласно таблице 4.1 из рабочих проб известной марки и концентрации основных питательных элементов (N, P, K) и S создавалась база данных физических

и химических свойств для оценки различных параметров исследуемых объектов. Так, каждый из исследуемых объектов представлен как минимум в 5 параллельных измерениях. Способы пробоподготовки, приведены в таблице 4.2.

Таблица 4.2. Стадии пробоподготовки.

Тип пробоподготовки	Истирание до < 500 мкм	Истирание до < 100 мкм	Сушка	Прессование
1	-	-	-	+
2	+	-	-	+
3	+	+	-	+
4	+	-	+	+
5	+	+	+	+

Пробы без прессования в большинстве случаев не использовались (если не оговорено обратное), поскольку даже для запрессованных объектов разных фракций наблюдалось изменение плотности и, как следствие, мертвого времени ЭД детектора. Для частичного нивелирования данного явления использовалось прессование, которое увеличивает время пробоподготовки, но, с другой стороны, обеспечивает возможность работы в вакууме и получения более достоверной информации о химическом составе пробы.

4.2.2 Общий алгоритм пробоподготовки рабочих проб (излучателей) к РФ анализу

Среднюю пробу сложных фосфорсодержащих минеральных удобрений (массой от 0,5 до 1,5 кг) помещали в чистую, сухую, плотно закрываемую стеклянную или полиэтиленовую тару. На тарную этикетку наносили следующие данные:

- наименование предприятия-изготовителя;
- наименование продукта, марку;
- номер партии;
- дату отбора пробы.

Среднюю пробу методом последовательного квартования или деления на делителе проб разбивали на аналитические пробы массой от 30 до 50 г.

Перед проведением анализа аналитическую пробу, при необходимости, просеивали через сита № 0,1 или 0,5 по ГОСТ 6613-86; остаток на сите растирали в агатовой или корундовой ступке до полного прохождения через сито. При необходимости 30 г. приготовленной пробы высушивали в сушильном шкафу при температуре (70±2)^{0}С до постоянной массы. Массу считали постоянной, если разность двух последовательных взвешиваний после сушки не превышала 0,0004 г. (согласно ГОСТ 5382-91). Затем чашку с содержимым охлаждали в эксикаторе по ГОСТ 25336-82 над хлористым кальцием или силикагель – индикатором по ГОСТ 8984-75. Высушенную пробу хранили в воздухонепроницаемом сосуде такой вместимости, чтобы проба полностью заполняла его. Допускается сушить часть аналитическую пробы массы порядка 5 г во влагомере до постоянной массы (расхождение на должно превышать 0,001 г в течение 5 минут).

После необходимых процедур изготавливался «излучатель» - запрессованная таблетка части аналитической пробы. Из исходной пробы случайным образом отбирали три навески массой около 0,1 г (на кончике шпателя) для прессования в таблетки. Прессование проводили в виде «сэндвич-структуры» в пресс-формах диаметром 20 мм. На первом этапе фигурным пуансоном прессовали тарелочку из борной кислоты массы порядка 3 г при усилии порядка 100 бар (3 т/см^2) в течении 30 секунд. Затем в полученную тарелочку помещали навеску пробы массой порядка 0,3 г так, чтобы проба не попадала на края тарелочки. Затем прессовали обычным пуансоном при 260 – 280 бар (10-11 т/см^2) в течении 30 – 60 секунд.

4.2.3 Общая процедура подготовки проб к анализу на физические свойства

4.2.3.1 Крупность объектов и степень обработки кондиционирующими добавками

Анализ объектов проводили по двум процедурам.

1. Для определения гранулометрического состава и степени обработки к.д. навески удобрений. Аналитическую пробу гранул делили на делителе проб до

массы порядка 5 - 10 г, затем полученную навеску помещали в специальную кювету и измеряли на оптическом стенде три раза с перемешиванием гранул.

2. Для определения крупности запрессованных объектов. Запрессованную пробу измеряли на оптическом стенде и рассчитывали значения крупности частиц по разработанному алгоритму с занесением полученной информации в базу данных исследуемых объектов.

4.2.3.2 Проводимость разбавленных растворов удобрений

Аналитическую навеску гранулированных удобрений последовательно делили на делителе проб до массы порядка 1 – 5 г и перемалывали на роторной ножевой мельнице в течении 15 минут. Из перемолотой фракции отбирали навеску массой 0,1 г. и растворяли в 100 г. дистиллированной воды для получения раствора концентрацией 0,1 масс. %. Раствор перемешивали на вибростенде в течении 30 мин. Полученный раствор фильтровали на фильтре «белая лента». Проводимость фильтрата измеряли с помощью кондуктометра.

4.3 Алгоритм проведения экспериментальной работы

4.3.1 Разработка установки для исследования объектов анализа.

На безе цифровой USB видеокамеры разработана макетная установка для получения информации по физическим свойствам излучателей (крупность частиц, гранулометрический состав, плотность поверхности излучателя и т.д.).

Макетная установка представляет из себя стенд из непрозрачного материала с веб-камерой ручной фокусировки, оснащенный системой углового освещения обычной и УФ-лампами. Используемое оборудование должно обладать характеристиками не хуже приведенных в таблице 4.3. Макетная установка подключается к персональному компьютеру по протоколу USB. Применяемое ПО дает возможность откалибровать систему пиксель/мм, сделать фотографию объекта и рассчитать по ней все интересующие физические параметры пробы.

Более подробно работа с установкой будет рассмотрена в главе 5 настоящей работы.

Таблица 4.3. Характеристики используемого оборудования.

Устройство	Параметр	Значение
фото- или видеокамера	разрешение	не менее 640x480
	фокусное расстояние	2,8 – 12 мм
	датчик	½. 7`` CMOS
УФ-лампа	длинна волны	≤ 370 нм
	световой поток	≥ 50 люменов
Лампа освещения	длины волн	> 400 нм
	световой поток	≥ 50 люменов

4.3.2 Создание программы накопления и обработки данных

В рамках проведенной работы создано и зарегистрировано ПО на языке Python 2.7 для обработки информации, получаемой от аппаратного рентгено-оптического комплекса, расчета концентраций и хранения полученных данных в базе данных исследованных объектов. Более подробно программное обеспечение будет рассмотрено в главе 7 настоящей работы. Программное обеспечение позволяет работать со всеми алгоритмами, использованными в ходе выполнения данной работы, часть которых приведена в приложении А.

4.3.3 Методика получения физических параметров проб с использованием оптической приставки

Созданные излучатели исследовались на качество поверхности с помощью макетного стенда с веб-камерой (пункт 4.2.3 настоящей работы). С использованием стенда записывали изображения поверхности объектов (гранул или прессованных таблеток), по которым рассчитывали необходимые признаки по алгоритмам, приведенным в пункте 3.2.1 настоящей работы. Оценка параметров проводилась 3 раза с поворотом пробы, результат усредняли. Данные каждого измерения заносили в базу данных собственного ПО и дополнительно сохраняли в *.csv файле.

4.3.4 Методика получения физико-химических параметров проб с использованием РФ-спектрометра

4.3.4.1 Подбор параметров работы РФ-спектрометра

Для каждого типа исследуемого объекта подбирали условия измерений на ЭД РФ-спектрометре, при которых минимизируются показатели зашумленности, количества неразделенных характеристических линий и т.д. В качестве настраиваемого параметра рассматривали время экспозиции, поскольку величина напряжения и тока ограничены показателем мертвого времени детектора, которое не должно превышать 20 % для обеспечения более точного контроля объектов различной плотности. Приборы производства АО «Научные приборы» оснащены системой корректировки времени измерения спектра пробы по мертвому времени детектора, что позволяет нивелировать небольшие изменения плотностей объектов.

4.3.4.2 Получение физико-химических параметров объектов

При подобранных оптимальных условиях измерений каждый объект анализировался в пяти параллелях с 3 поворотами кюветы. По записанным спектрам вычислялись физико-химические параметры объекта, согласно пункту 3.2.2 настоящей работы. Выделенные признаки заносились в базу данных.

По выделенным признакам рассчитывали модели классификации для определения физических параметров, таких как:

- максимальная фракция объекта;
- наличие предварительной сушки.

Так же строились и оптимизировались модели регрессии для определения марки удобрений и концентраций химических элементов, входящих в их состав. Исследовались процессы перехода с выпуска одной марки удобрений на другую и проводилась кластеризация с понижением размерности для экспрессного мониторинга качества производимой продукции и расчета оптимального пути перехода на примере марки NP(S) 12-40(10) и NP(S)+S+Zn 12-40(6)+3+1.

5 Модернизация оборудования и выделение физико-химических признаков объектов

Одним из ключевых результатов проведенной работы является создание усовершенствованного и более информативного программно-аппаратного комплекса на базе ЭД РФ спектрометра «РЕАН», производства АО «Научные приборы» и оптического анализатора собственной разработки. Предложенная схема позволяет проводить быстрый и неразрушающий анализ промышленных объектов. При этом в качестве спектрометра может использоваться как стационарная, так и переносная версия ЭД РФА. Предложенная оптическая схема позволяет определить такие показатели, как: гранулометрический состав и степень обработки кондиционирующей добавкой, которая является наиболее дорогим реагентом в производстве. Данные параметры особенно актуальны при их получении в режиме on-line – в течении 15 минут от отбора пробы, что и позволяет сделать разработанный программно-аппаратный комплекс.

5.1 Макетный образец оптического анализатора

Для сбора информации о физических свойствах исследуемых объектов и их качества, в том числе и качества пробоподготовки, создан лабораторный стенд автоматического оптического контроля (рисунок 5.1 и 5.2).

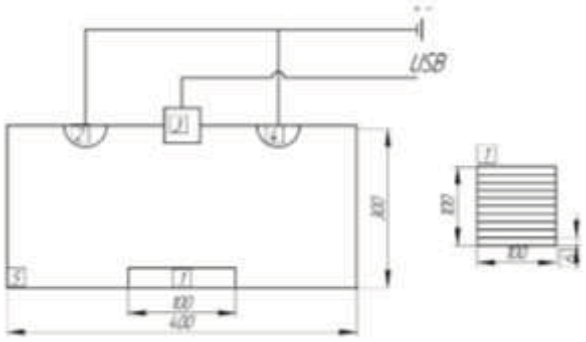


Рисунок 5.1. Схема стенда для измерения качества поверхности излучателя (размеры в мм). 1 – кювета для работы с гранулами. 2 – лампа УФ освещения, 3 – веб-камера, 4 – лампа светодиодного освещения, 5 – поверхность не пропускающая внешнего света, 6 – рифленая поверхность кюветы для представительного анализа гранул.

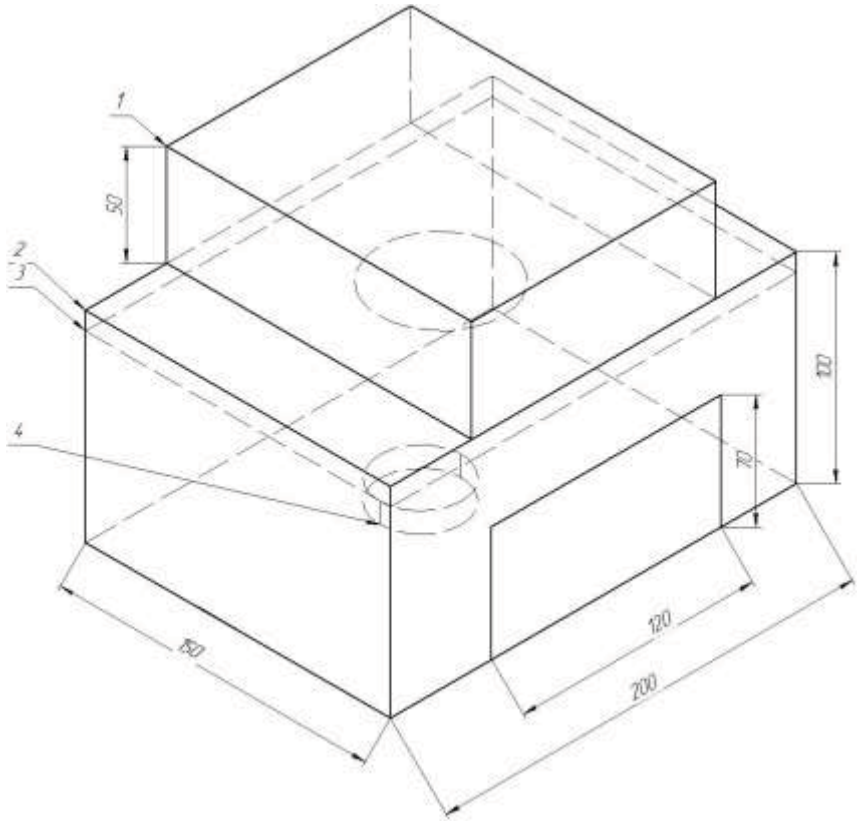


Рисунок 5.2. Трехмерная модель оптического регистратора. 1 – регистрирующее устройство, 2 – корпус, 3 – лента светодиодного освещения, 4 – исследуемый объект. Размеры приведены в мм.

Достоинствами данного прибора являются:

- малые габариты и вес;
- быстрая пробоподготовка;
- простая автоматизация измерений;
- простота и экономичность конструкции.

Данная установка может работать как с гранулами минеральных удобрений, так и с запрессованными излучателями для РФ анализа, обеспечивая получение информации о таких показателях качества как:

- гранулометрический состав (величина, контролируемая по ГОСТ для каждого типа промышленно выпускаемого минерального удобрения);
- степень обработки кондиционирующими добавками (самый дорогой реактив на производстве);

- максимальная фракция прессованного порошка (значимо влияет на качество РФ анализа).

При работе с гранулами используется специально разработанная кювета с рифленой поверхностью. Величина шага поверхности кюветы равна среднему диаметру исследуемых гранул для обеспечения распределения гранул и их фиксации по длинной оси для более представительного анализа гранулометрического состава. Данные по средней фракции приведены в таблице 5.1.

Таблица 5.1. Средняя фракция исследуемых гранул удобрений.

Тип удобрения	Диапазон гран. состава по ГОСТ, мм	Гранул в диапазоне, масс. %	Среднее значение*, мм
NPK	1 – 6	> 97	2,9
NPK(S)	1 – 6	> 97	3,0
NP(S)	2 – 5	> 90	3,2
Аммофос (МАФ)	1 – 6	> 97	3,5
ДАФ	1 – 6	> 97	3,5

* установлено экспериментально по навеске массой 200 г

Предложенная конструкция обеспечивает быстрый и удобный способ получения различной физической информации об исследуемых пробах.

5.2 Выделение физических параметров проб с использованием оптического регистратора

Разработанный оптический анализатор позволяет исследовать поверхность образцов для РФ-анализа при различной пробоподготовке (рисунок 5.3).

а)

б)

в)

Рисунок 5.3. Пример изображения излучателя а) – гранулы МАФ 12-52, б) – фракция < 500 мкм NPK(S) 4-30-15(16), в) – фракция < 100 мкм NPK 16-16-8.

На следующем этапе выделяется область поверхности излучателя с разрешением не менее 100×100 пикселей (рисунок 5.4).

Рисунок 5.4. Выделенная область поверхности излучателя удобрения NPK(S) 4-30-15(16) фракцией не более 500 мкм.

Полученное изображение переводится в черно-белый формат (в градациях серого), дифференцируется для устранения тренда освещения и сглаживается для нивелирования шумов согласно пункту 3.2.1 настоящей работы. По полученному изображению строится карта поверхности (рисунок 5.5).

По карте поверхности выделяются максимумы и минимумы яркости (аномалии на карте) и рассчитывается их удельная площадь и количество, согласно главе 3 настоящей работы и затем полученные значения заносятся в базу данных: матрица «объекты-признаки».

Для каждой из описанных выше стадий проведена процедура оптимизации параметров для расчета карты поверхности и дальнейшей классификации по выделенным

признакам. Для этого проанализировано порядка 150 проб удобрений (таблица 4.1) при различных стадиях пробоподготовки: запрессованные гранулы, порошок фракции < 500 мкм и порошок фракции < 100 мкм. Для сравнения дополнительно исследованы гранулы удобрений без предварительного прессования.

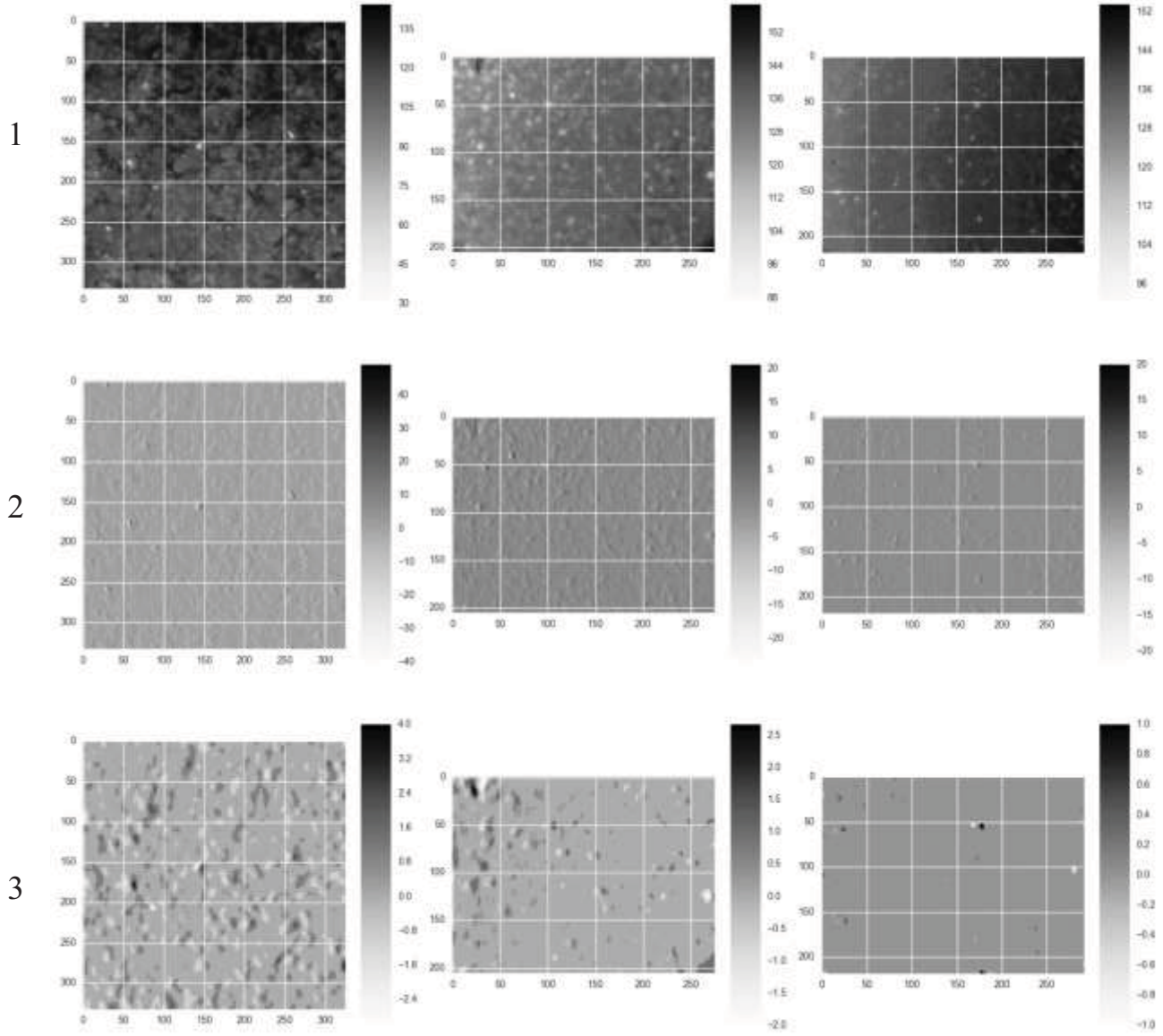


Рисунок 5.5. Пример работы алгоритма а) – запрессованные гранулы, б) – запрессованный порошок < 500 мкм, в) – запрессованный порошок < 100 мкм. На изображениях 1 ряда показана исходная поверхность черно-белого изображения, на 2 ряду – поверхность после дифференцирования, на 3 ряду – карта поверхности после сглаживания медианным фильтром.

В качестве метрики качества работы алгоритма выбран коэффициент корреляции с фракцией, которая является одной из целевых переменных и наименее зависима от химического состава пробы (информация, доступная из РФ спектров) по сравнению с типом и маркой удобрения. Поскольку выделяемые признаки представляют собой вещественные значения и по предварительной оценке линейно связаны с искомым параметром – классификацией по крупности частиц, использовали коэффициенты корреляции Пирсона и Спирмана. Коэффициент Пирсона оценивает непосредственно линейную корреляцию между двумя переменными, тогда как коэффициент Спирмана является ранговым параметром и оценивает силу монотонной взаимосвязи и добавлен для более общей оценки взаимосвязи.

Предварительная оценка карты линейных корреляций проводилась для начального приближения 10 и 1 для окна медианного фильтра и константы определения контуров. Использованные признаки приведены в таблице 5.2

Таблица 5.2. Использованные оптические признаки пробы

Признак	Описание
тип	тип удобрения
фракция	фракция пробы
площадь включений	удельная площадь включений, определенная по карте поверхности
количество включений	удельное количество включений, определенное по карте поверхности
средняя яркость	средняя яркость изображения в градациях серого

Для первоначальной оценки признаков по их значимости построена карта корреляций (рисунок 5.6). На данном графике приведены коэффициенты корреляции Пирсона по каждой паре признаков. Чем ближе по модулю коэффициент корреляции в единице, тем сильнее линейная связь между признаками.

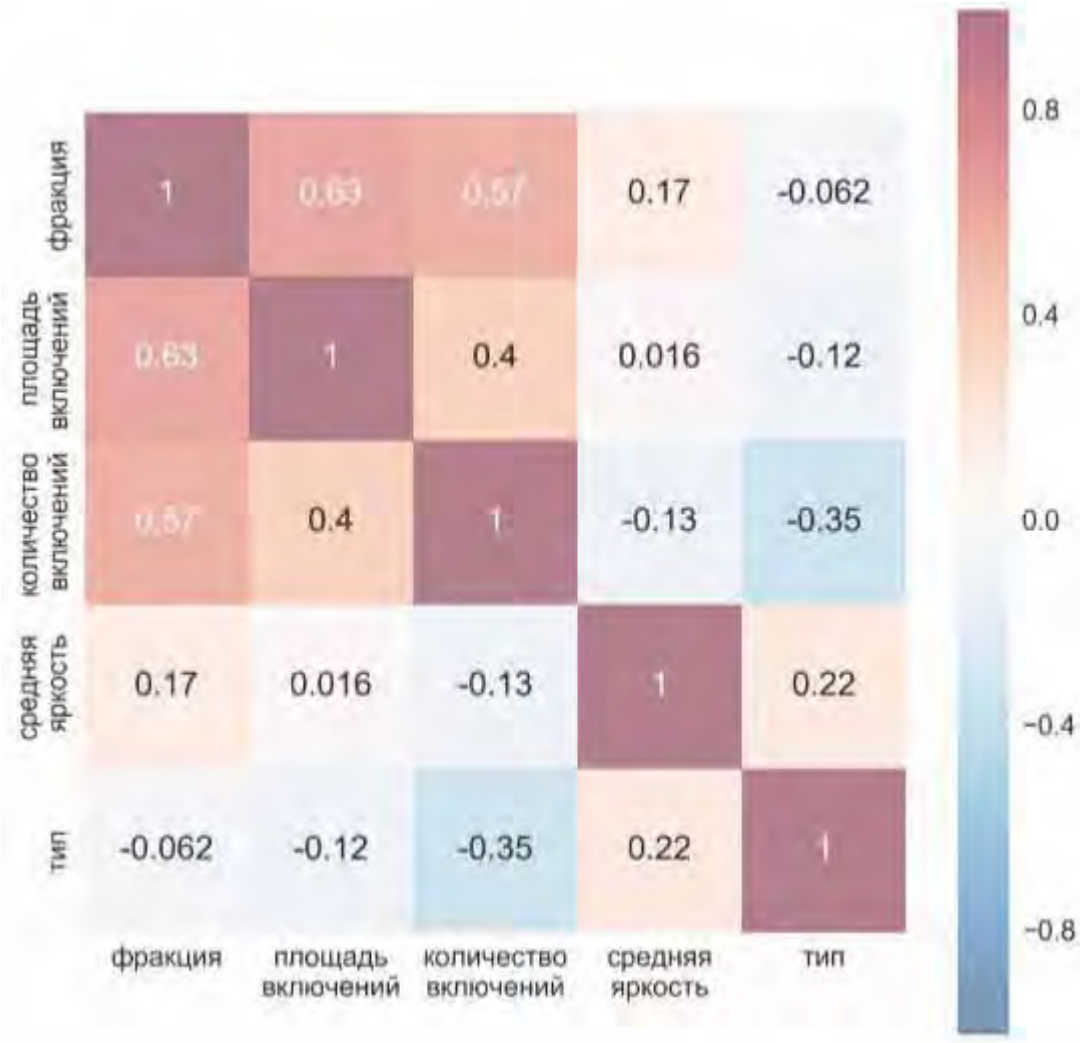
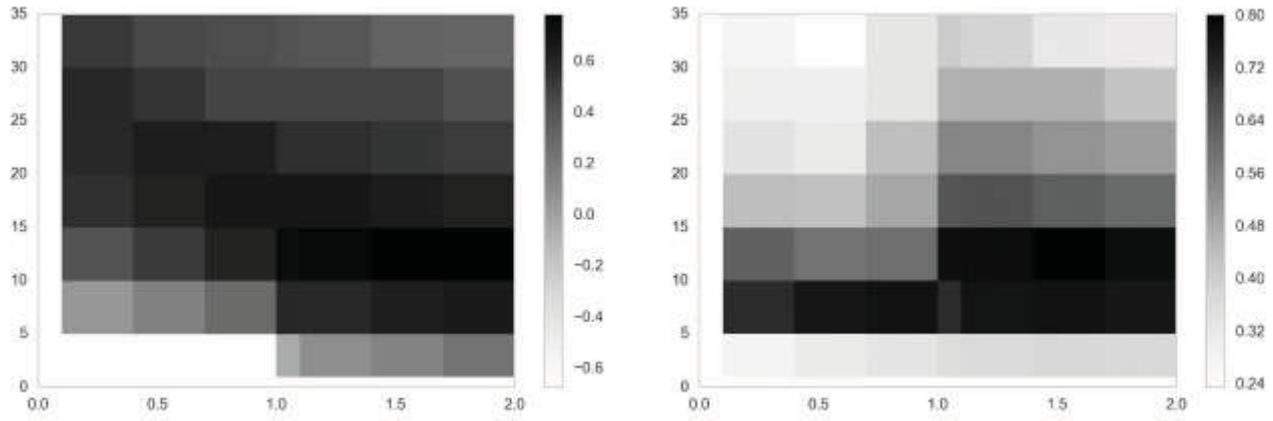


Рисунок 5.6. Карта линейных корреляций между оптическими признаками пробы.

Исходя из полученных данных подтвердилось предположение о линейной корреляции между фракцией и выделенными признаками. Для оптимизации данного параметра использовался перебор по сетке. В случае медианы использовался диапазон [1; 31] с шагом 5, использовать большие значения окна не целесообразно, поскольку длина и ширина изображения составляют 100 пикселей. Для константы контуров выбран диапазон [0,1; 2,0] с шагом 0,1, как наиболее универсальный.

Для поиска оптимальных параметров построена трехмерная карта корреляции удельной площади и количества определенных контуров с фракцией по Спирману для всех возможных значений окна медианного фильтра и константы определения контуров (рисунки 5.7 и 5.8).



а) б)

Рисунок 5.7. Карта изменения корреляции Спирмана в зависимости от оптимизационных параметров а) - корреляция по удельному количеству аномалий, б) – корреляция по удельному размеру аномалий

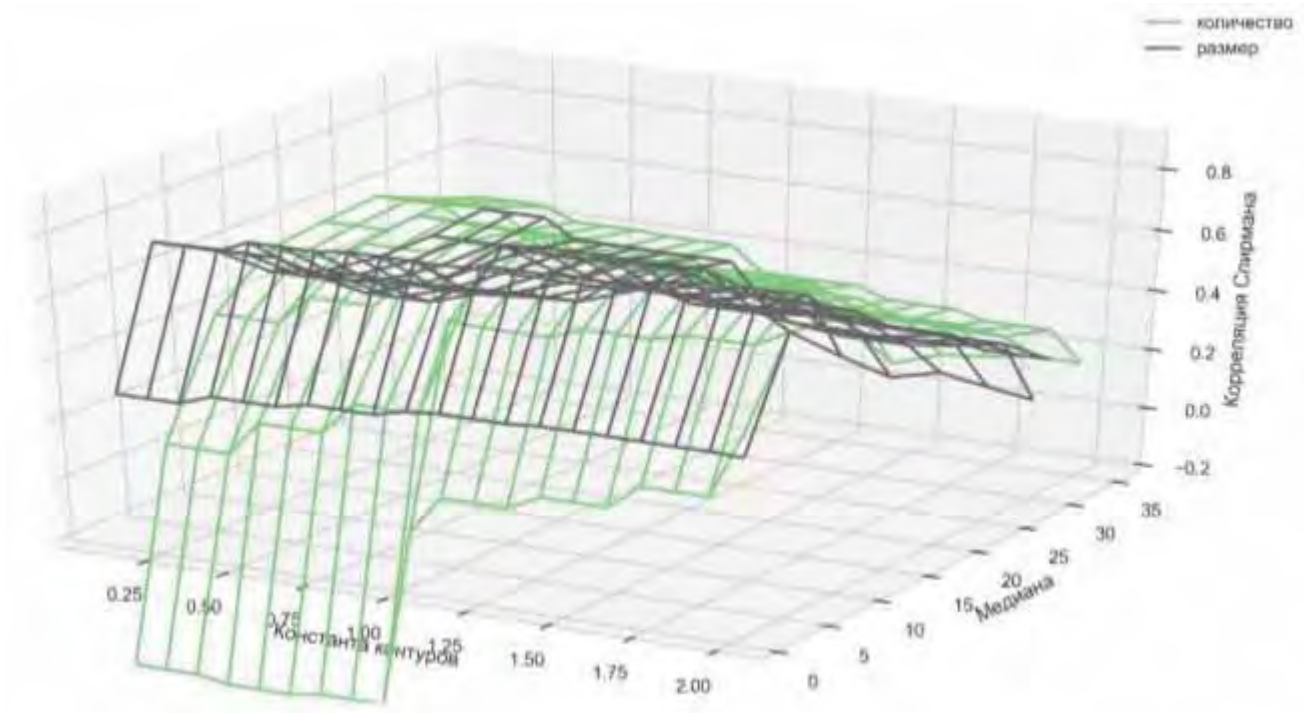


Рисунок 5.8. Трехмерное изображение корреляции по Спирману в зависимости от оптимизируемых параметров

Дополнительно представлена трехмерная карта корреляции по Пирсону (рисунок 5.9).

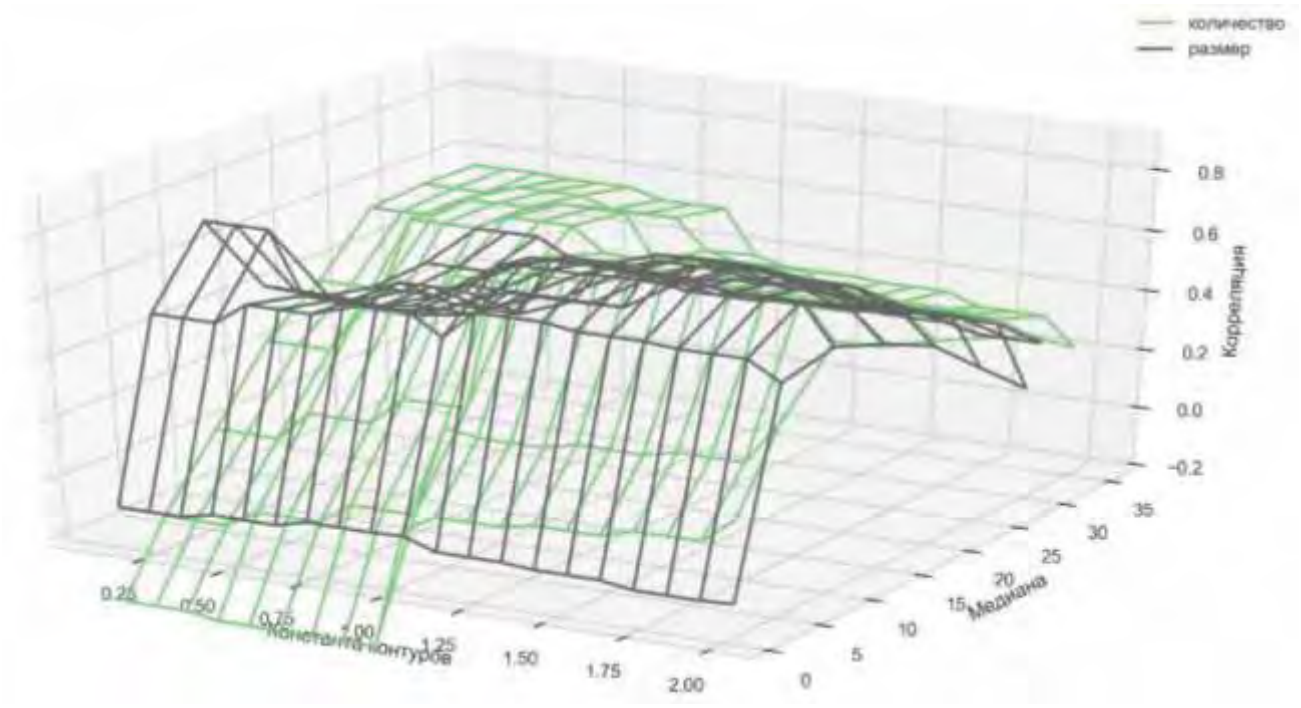


Рисунок 5.9. Трехмерное изображение корреляции по Пирсону в зависимости от оптимизируемых параметров

По предварительной оценке оптимальными параметрами являются окно медианного фильтра в 15 пикселей и константа поиска контуров в 1,3 единицы, когда обе корреляции по удельному размеру и удельному количеству контуров принимают высокие значения. Для оптимизации будем рассматривать корреляцию Спирмана как более универсальную. Интересно заметить, что корреляция Спирмана по удельной площади контуров хорошо работает в диапазоне 5 – 10 и 0,4 – 1,0 или 10 – 15 и 1,0 – 2,0 пикселей медианного фильтра и единиц константы контуров соответственно. Однако корреляция по количеству включений лучше всего работает при более грубых параметрах 10 – 15 и 1,0 – 2,0. Уточним выбранные параметры. Изменение корреляции в зависимости от окна медианного фильтра приведено на рисунке 5.10.

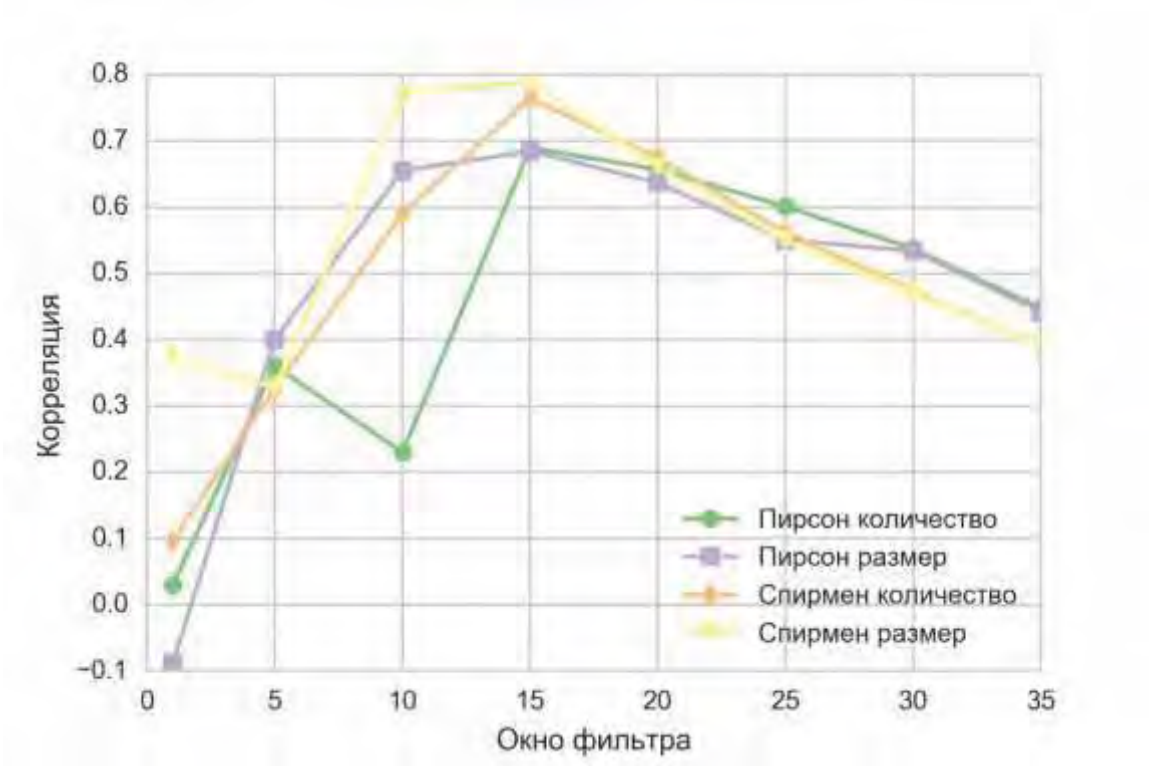


Рисунок 5.10. Зависимость коэффициентов корреляции от параметра окна медианного фильтра.

Результатом оптимизации является максимум корреляции при окне в 15 пикселей – корреляция по Спирману достигает 80 %. При изменении окна фильтра с 0 до 5 пикселей наблюдается уменьшение корреляции удельной площади аномалий по Спирману и увеличение корреляции удельного количества аномалий по Пирсону с фракцией объекта. Обнаруженное явление является аномальным, поскольку корреляция Пирсона (сила линейной взаимосвязи) по сути является частным случаем корреляции по Спирману (сила монотонной взаимосвязи). Данный эффект скорее всего вызван наличием сильной шумовой компонентой в данных.

Далее, при выбранном параметре окна сглаживания, проведен поиск коэффициента определения контуров. Зависимость корреляции фракции и выделенных из карты поверхности признаков от константы контуров приведена на рисунке 5.11.

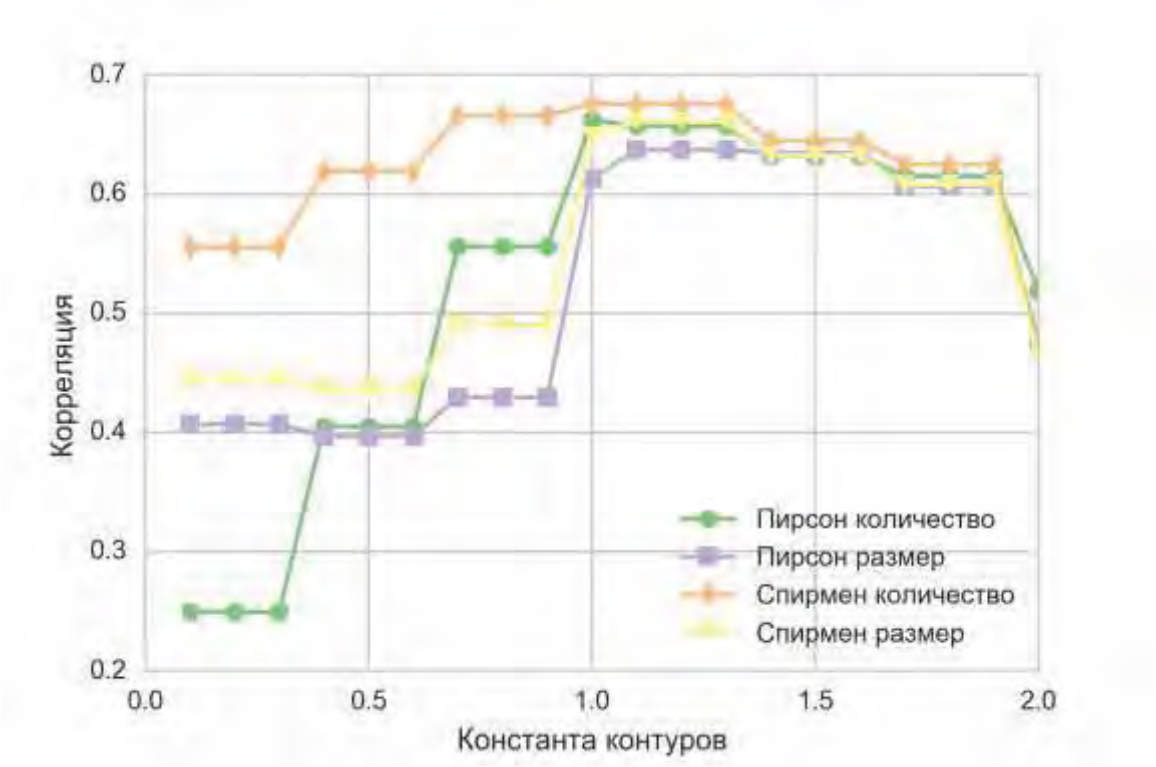


Рисунок 5.11. Зависимость коэффициентов корреляции от параметра аппроксимации контуров.

Оптимальным параметром является начальное приближении в 1,1 – 1,4 единицы с наиболее выраженной корреляцией. Интересным является приближении качества классификации по размеру и количеству аномалий при росте коэффициента. Данное явление скорее всего обусловлено загрублением алгоритма поиска контуров. Так же стоит отметить определенную дискретную структуру работы алгоритма. Подобное поведение вполне ожидаемо, поскольку контуры на изображение конечны и расположены в определенных областях. Изменение константы контуров по сути уменьшает шумовые составляющие (контуры малого размера), но не изменяет поведения алгоритма.

Дополнительно изучено поведение самих определяемых признаков (размера и количества контуров) в зависимости от оптимизационных параметров (рисунок 5.12).

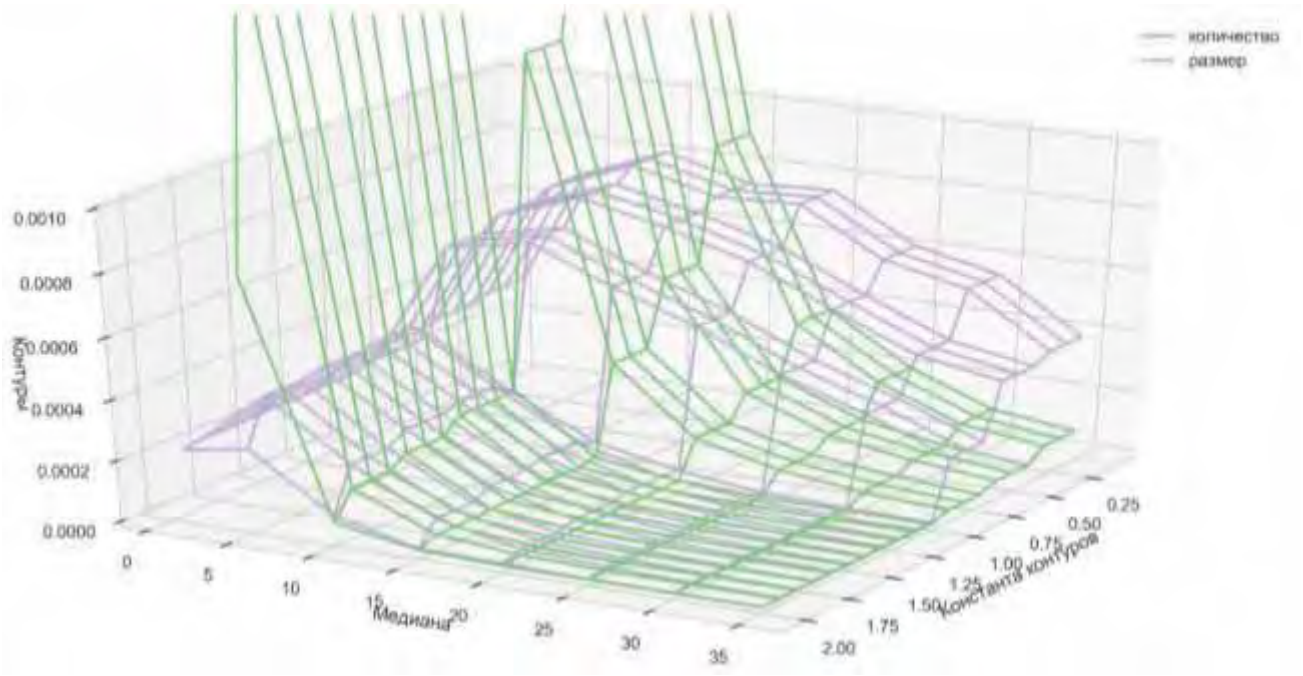
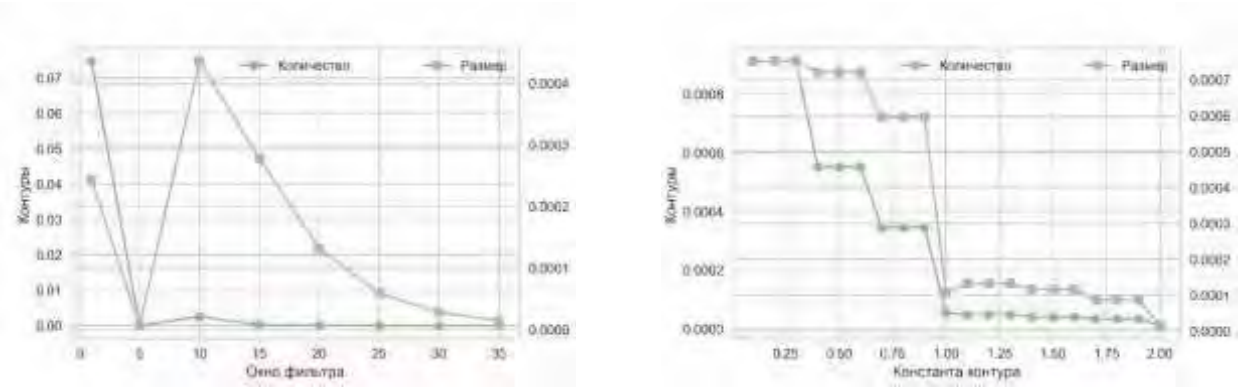


Рисунок 5.12. Трехмерное представление среднего удельного количества и среднего удельного размера определенных контуров в зависимости от оптимизируемых параметров

Интересно отметить, что при выбранных ранее оптимальных параметрах наблюдается выравнивание среднего удельного количества и средней удельной площади найденных контуров. С учетом стабилизации работы алгоритма определения контуров при константе в 1,3 – 1,5, выбранные параметры в 15 пикселей и 1,3 единицы для окна медианного фильтра и константы контуров представляются нам оптимальными.

Двухмерные представления для оптимальных параметров приведены на рисунке 5.13. Медианный фильтр с окном, менее 10 пикселей ведет себя нестабильно, что уже отмечалось выше при оценке корреляции. Выявленное поведение метрики качества скорее всего обусловлено наличием сильных шумовых составляющих в данных. После отметки в 10 пикселей медианный фильтр ведет себя закономерно, уменьшая количество контуров с ростом величины окна сглаживания. В то же время константа контуров резко сокращает количество (или средний размер) контуров к 1,0 и далее стабилизируется.



а) б)

Рисунок 5.13. Зависимость удельного количества определенных контуров от окна медианного фильтра (а) и удельного количества определенных контуров от константы контуров (б).

Данное поведение подозрительно и возможно вызвано плохим распознаванием контуров на «хороших» пробах – фракцией 100 и 500 мкм. Проверим данное предположение для фракции 500 мкм – построим карты бинарных корреляций (рисунок 5.14).

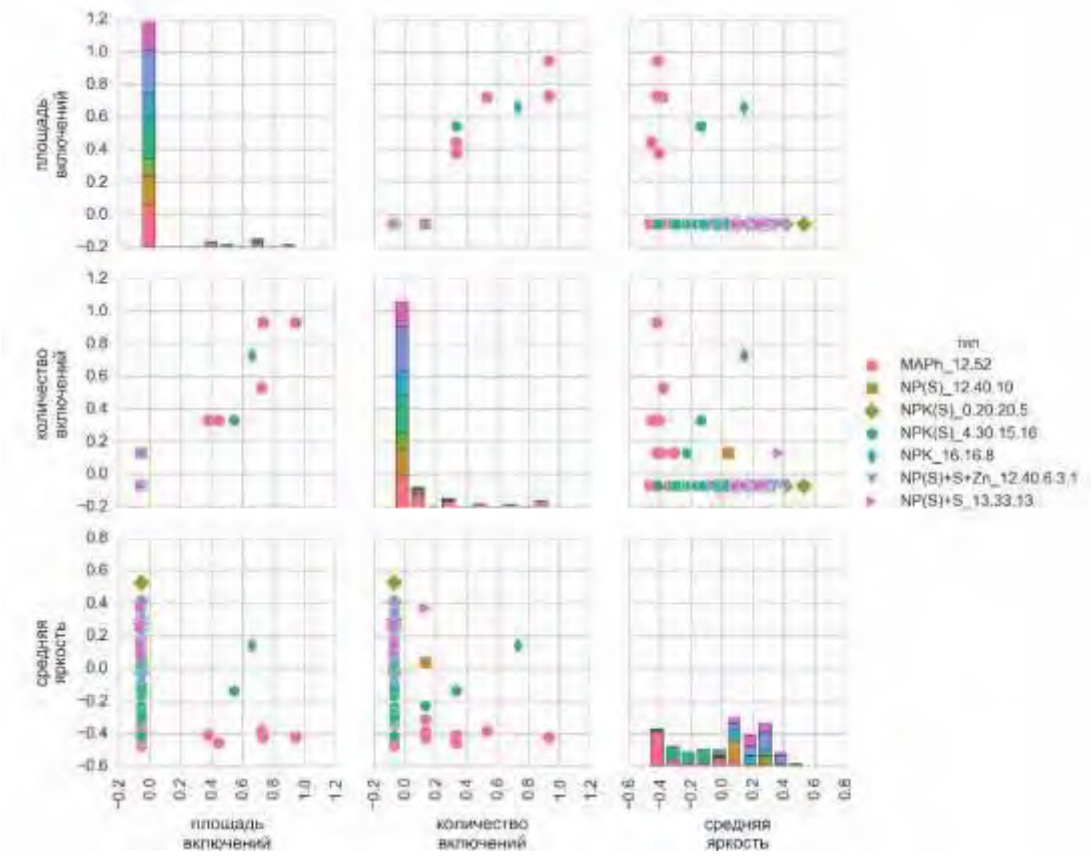


Рисунок 5.14. Карта бинарных корреляций для фракции 500 мкм при параметрах алгоритмов 15 пикселей и 1,5

Действительно, параметры в большинстве случаев обнуляются. При этом увеличение коэффициента корреляции вызвано сокращением количества классов. Таким образом, константа контуров не должна 1 и оптимальными значениями являются 10 пикселей и 0,7 – 0,9. Аналогично, проверим значения параметров найденных контуров при новых условиях для фракции в 100 мкм (рисунок 5.15).

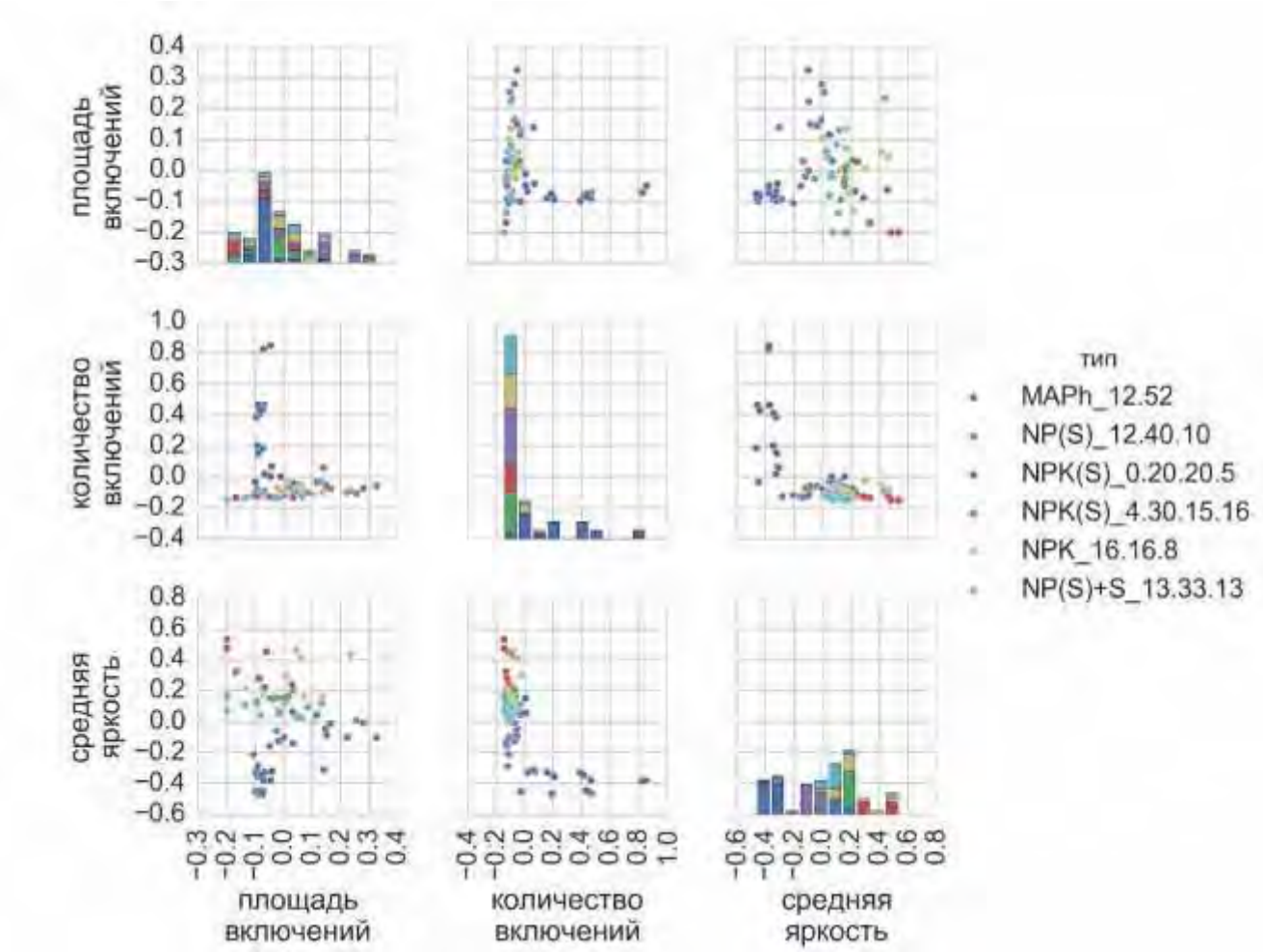


Рисунок 5.15. Карта бинарных корреляций для фракции 100 мкм при параметрах алгоритмов в 10 пикселей и 0,7

Обнуления параметров не происходит. Оптимальные параметры установлены. Полная карта линейных корреляций по Пирсону для выбранных параметров приведена на рисунке 5.16.

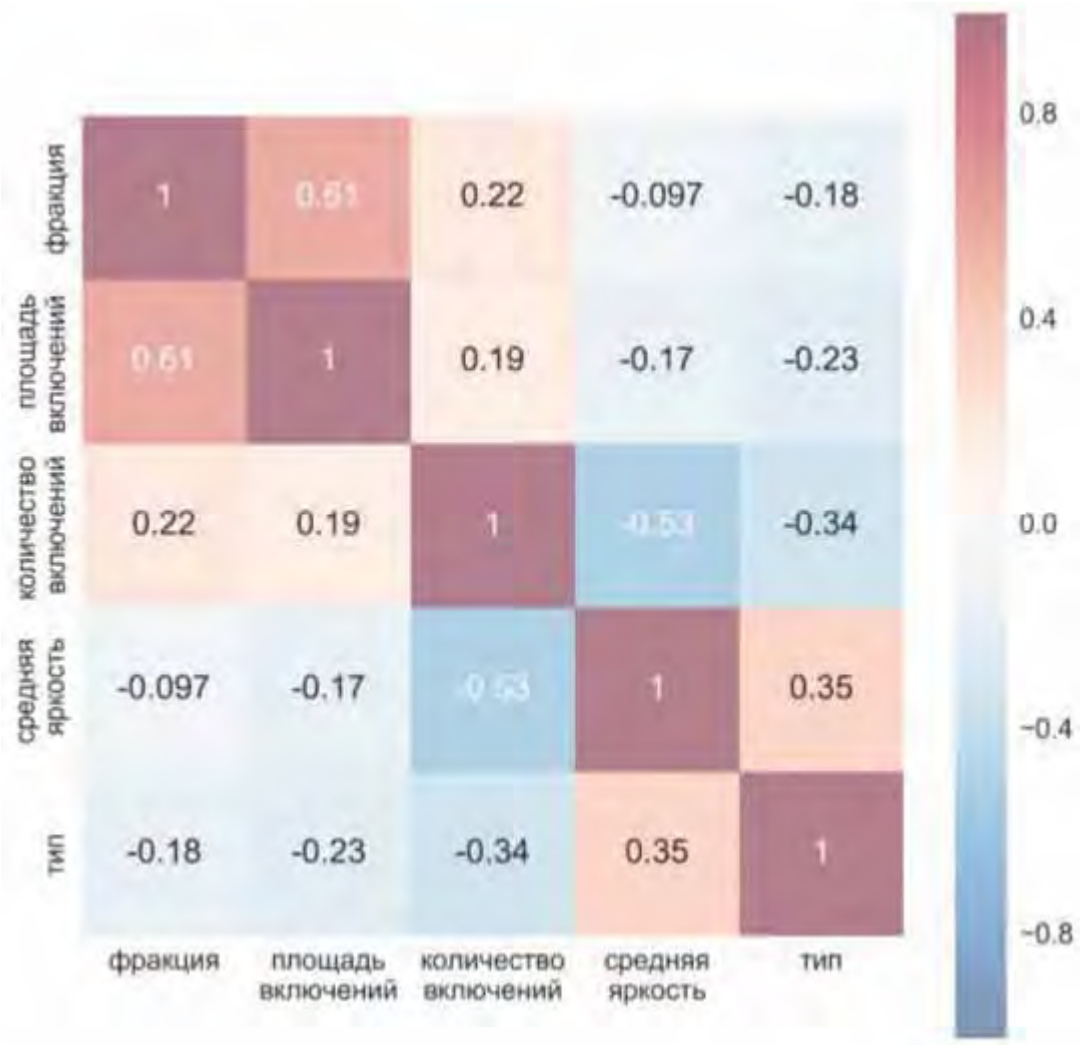


Рисунок 5.16. Карта линейных корреляций при окне в 15 пикселей и константе определения контуров в 1,3

При подборе параметров наблюдается достаточно сильная линейная корреляция с фракцией. Далее проведено выделение признаков и построена матрица «объекты-признаки» размером порядка 180×3. Пример выделения карты поверхности приведен на рисунке 5.17.

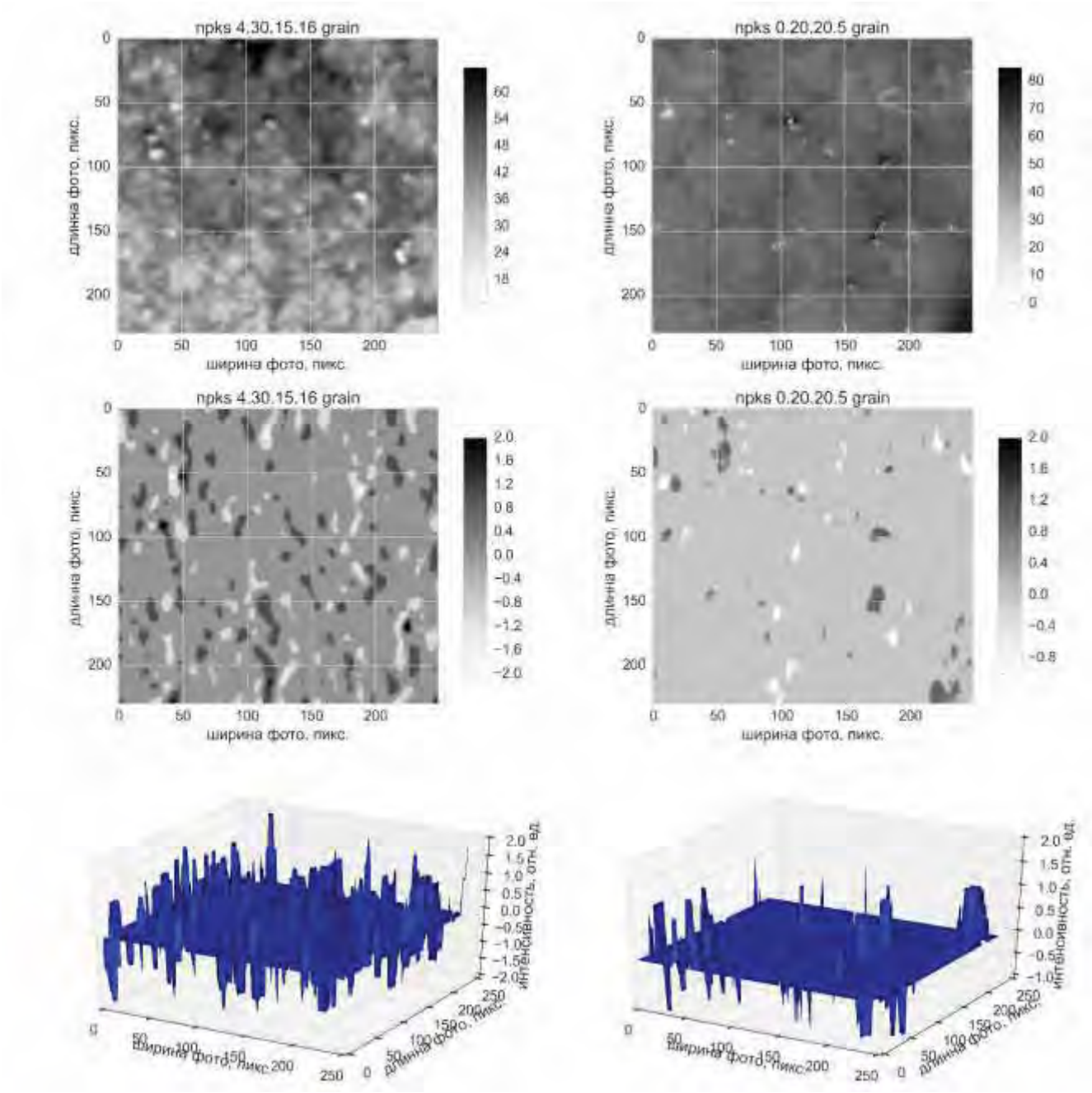
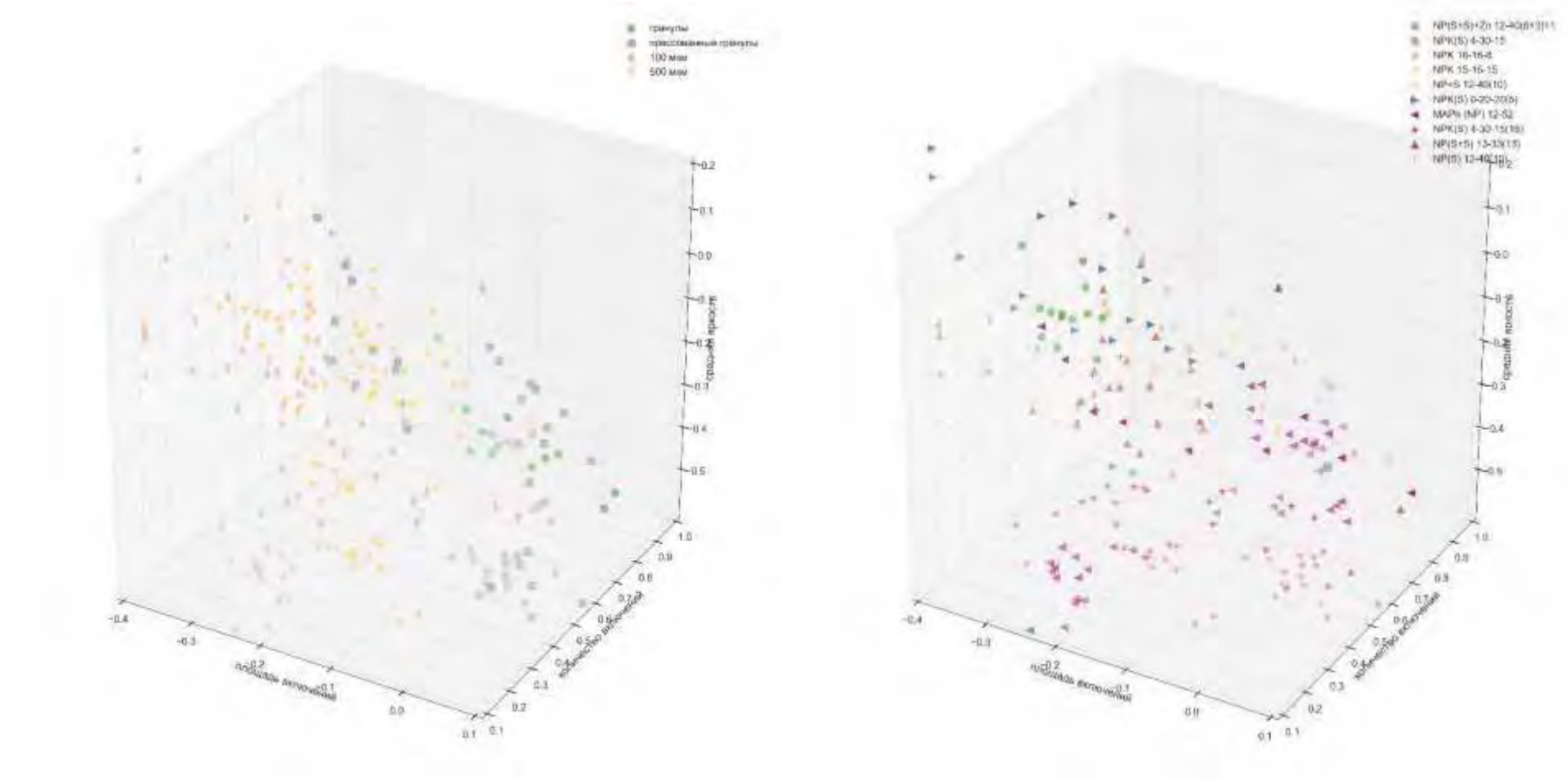


Рисунок 5.17. Выделение карты поверхности из фотографии объектов исследования.

Картой поверхности является основание трехмерного изображения, по которому определяются удельный размер и площадь аномалий (контуров). В связи с малым количеством признаков построено облако точек для фракции и типа исследуемого объекта и проведена классификация по алгоритму k-means (рисунок 5.18).



а) б)

Рисунок 5.18. Представление фракции (а) и типа (б) исследуемых объектов.

Классификация по фракции исследуемого объекта проводиться достаточно точно (точность и F-мера 81 % по алгоритму k-means). В целом облака точек частично разделимы в пространстве, чего нельзя сказать о классификации по типу объектов. Для обеспечения точной и воспроизводимой классификации исследуемых объектов по типу требуется выделение не только физических признаков, но и химических признаков из спектра.

5.3 Выделение физико-химических параметров из РФ-спектра проб

В рамках данного раздела проводится подбор универсальных условий измерения спектров всех исследуемых объектов. Подбор оптимальных параметров предварительной обработки спектров и выделения аналитических сигналов проводится с последующим составлением матрицы «объекты-признаки» по всем свойствам, полученным с помощью ЭД РФА.

5.3.1 Определение оптимальных условий записи спектров

На первом этапе, при расчете физико-химических параметров проб из РФ-спектра, необходимо подобрать оптимальные условия измерений спектров. Так, мертвое время детектора для каждого из исследуемых объектов не должно превышать 20 %, обеспечивая достаточную чувствительность, селективность и разрешающую способность спектрометра. Полезной, в данном случае, является функция корректировки времени измерения спектра в зависимости от значения мертвого времени детектора. Спектрометры «РЕАН» позволяют проводить такую корректировку. Поэтому, варьирование мертвого времени в пределах 20 % не должно значимо сказаться на качестве работы методов АБД. Так, с использованием РФ спектрометра «РЕАН» с 50 Вт молибденовой трубкой, проанализированы различные марки удобрений при напряжениях 10 – 40 кВ и током в 50 мкА. Определено при каком напряжении не происходит превышение 20% порога мертвого времени детектора. Установлено, что максимальным напряжением (с

возможностью варьирования тока) является напряжение в 25 кВ. Использовать напряжение менее 20 кВ не имеет смысла, поскольку теряется информация о когерентном и некогерентном рассеянии излучения трубки (Kα Mo = 17,4 кэВ). Также частично теряется информация об относительно тяжелых элементах, входящих в состав удобрений (например, стронций: Kα Sr = 14,1). Максимально возможным током для данного напряжения, без превышения порога мертвого времени в 20%, является 100 мкА. При этом не происходит явного перекрывания характеристических линий основных питательных элементов. Остается подобрать время экспозиции. Для этого исследован следующий набор проб (таблица 5.3):

Таблица 5.3. Рабочие пробы для оптимизации экспозиции

Тип пробы	**Разбавление**	**Диапазон экспозиции, с**
Апатит	-	[25,100] с шагом 25
	50 масс. % H_3BO_3	
NPK(S) 4-30-15(16)	-	
	50 масс. % H_3BO_3	

Общий вид полученных спектров приведен на рисунке 5.19. По полученным спектрам не выявлено значимой зависимости соотношения «сигнал – шум», а также дисперсии сигнала от времени измерения. В таблице 5.4 приведены усредненные значения соотношения сигнал/шум (по пяти параллельным измерениям).

Таблица 5.4. Зависимость сигнал/шум для некоторых признаков спектра (фосфор и молибден) от времени экспозиции.

Объект	**Экспозиция**	**Разбавление H_3BO_3**		**Без разбавления**	
		Среднее соотношение сигнал – шум, P	**Среднее соотношение сигнал – шум, когерентное рассеяние Mo**	**Среднее соотношение сигнал – шум, P**	**Среднее соотношение сигнал – шум, когерентное рассеяние Mo**
Апатит	25	11,49	6,31	10,44	11,05
	50	11,34	6,31	10,38	10,44
	75	11,30	6,25	12,08	10,88
	100	11,92	6,27	12,95	10,89
NPK(S) удобрение марки 4-30-15(16)	25	18,67	2,17	25,79	3,32
	50	17,62	2,16	24,29	3,33
	75	17,91	2,17	25,33	3,31
	100	17,63	2,19	25,07	3,31

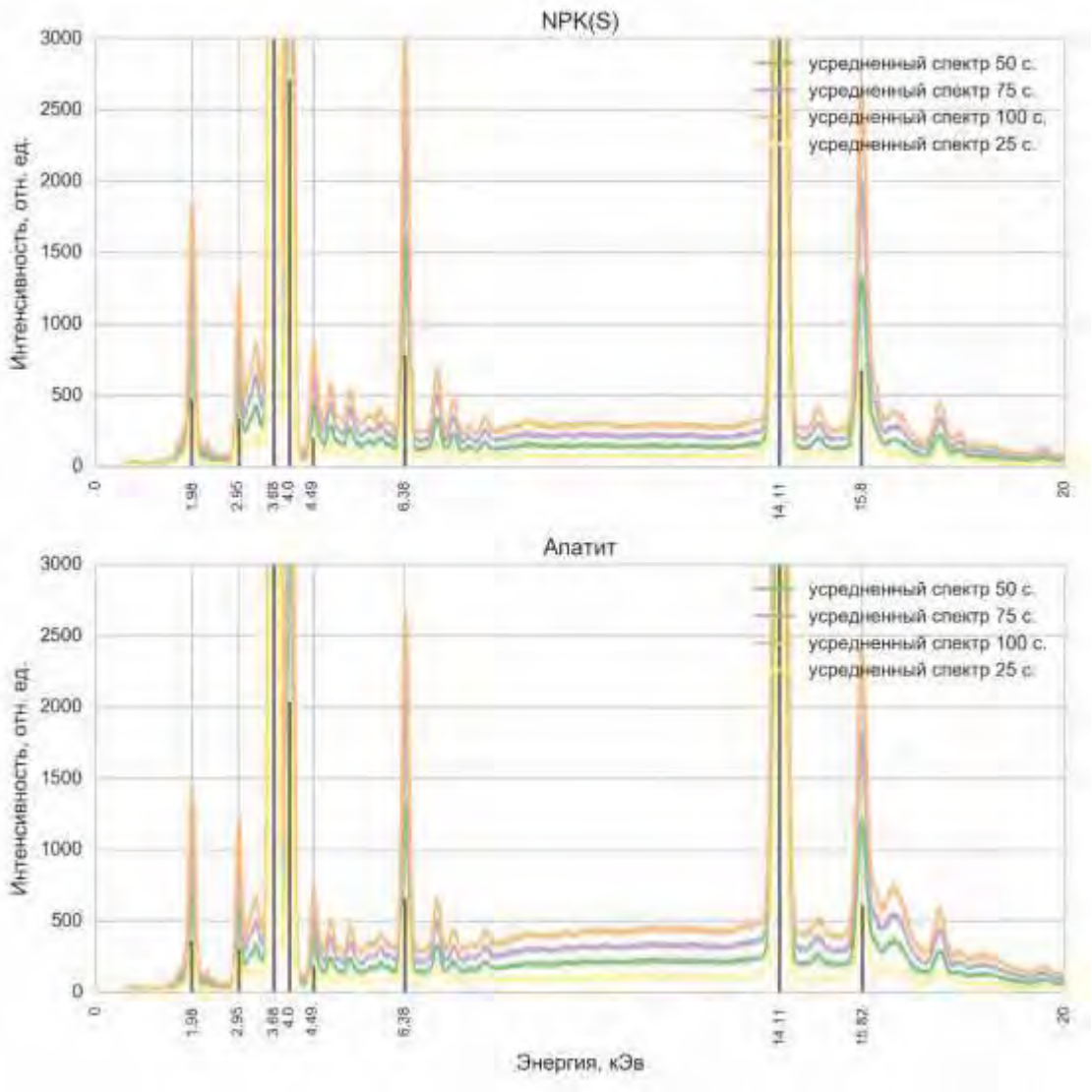


Рисунок 5.19. Спектры объектов, оптимизируемых по экспозиции

В приведенных данных не наблюдается структуры, а вариации значений скорее всего вызваны случайными шумами. Можно обратить внимание, что исследованные параметры зависят от мертвого времени детектора, а значит, и от плотности объекта. Таким образом, экспозиция в 50 секунд представляется оптимальным выбором. Проверим предположение, рассмотрев значимо ли изменяется дисперсия в каждом энергетическом канале спектра в зависимости от времени экспозиции. Рассчитанные значения приведены на рисунке 5.20.

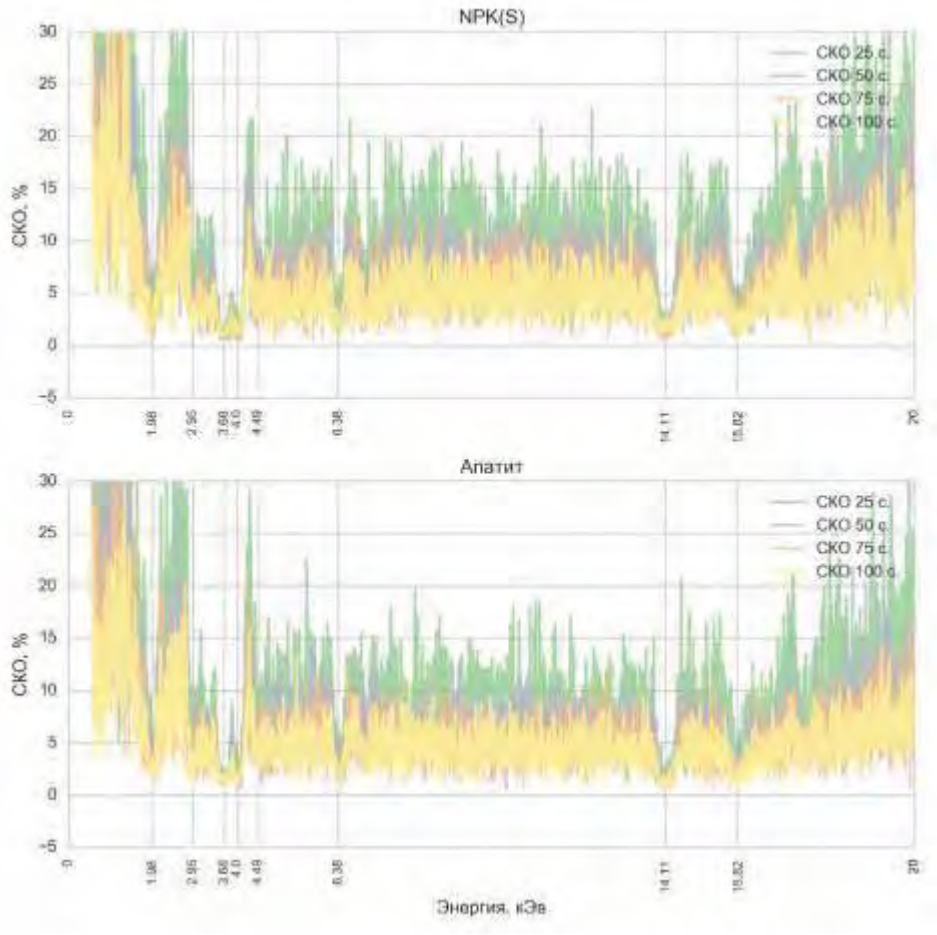


Рисунок 5.20. Зависимость относительного значения дисперсии каждого канала спектра от времени экспозиции.

Хотя значение дисперсии шумовой компоненты спектра сильно зависит от времени накопления сигнала, но для характеристических линий изменения не значимы (менее 5 % для любой экспозиции). Таким образом, на наш взгляд, лучше сократить время анализа и провести предварительное сглаживание спектра. Так, дисперсия спектров, сглаженных алгоритмом Савицкого-Голея с параметрами 3 и 11 (степень полинома и размер окна сглаживания соответственно) приведены на рисунке 5.21.

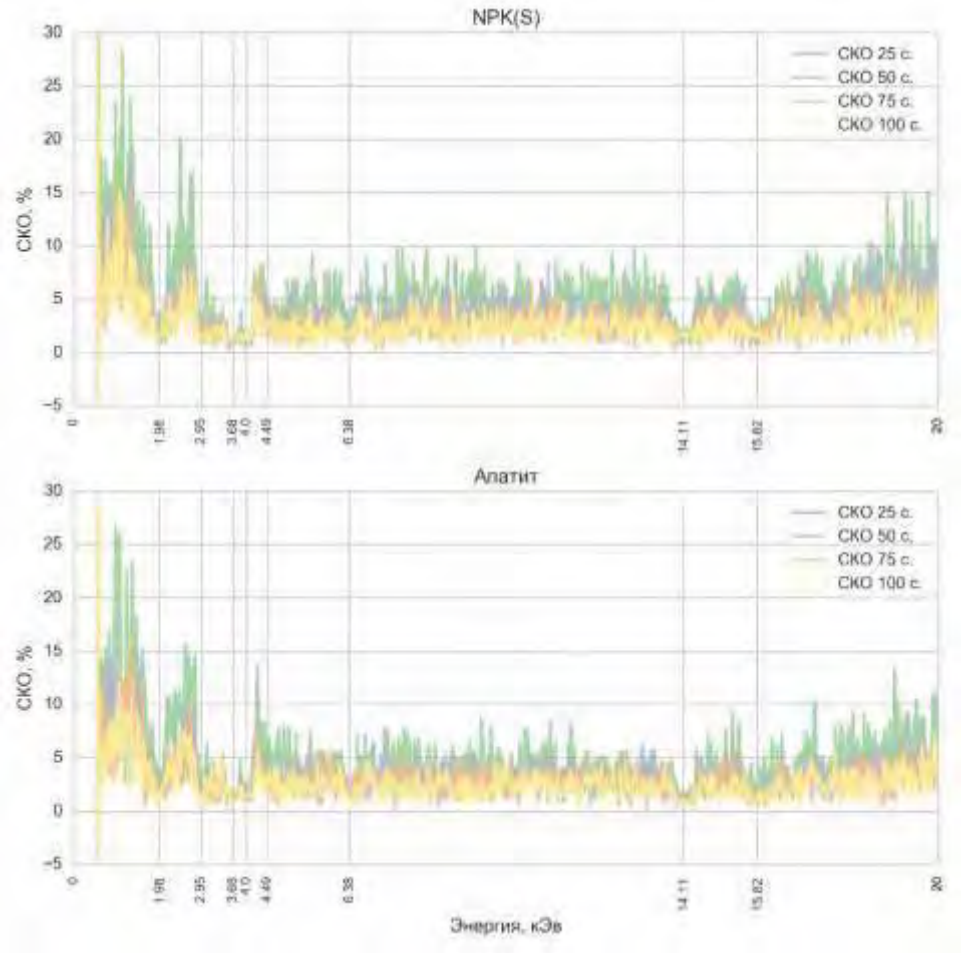


Рисунок 5.21. Зависимость относительного значения дисперсии каждого канала спектра от времени экспозиции после сглаживания.

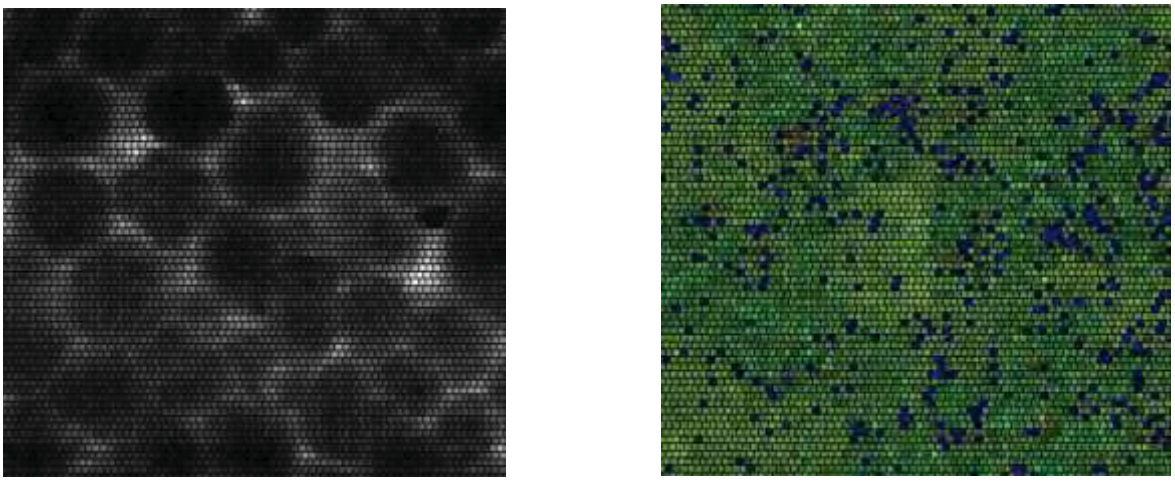
За исключением области легких элементов (менее 1 кЭв) дисперсии экспозиций 50 и 100 секунд очень похожи. Учитывая, что основные определяемые элементы находятся в области энергий более 1,5 кЭв, выбраны оптимальные условия записи спектров исследуемых объектов: 25 кэВ, 100 мкА и 50 с.

5.3.2 Оптимизация алгоритмов выделения физико-химических параметров из РФ-спектров

Каждый записанный спектр объекта проходил предварительную предобработку согласно пункту 3.2.2 настоящей работы. Для каждого алгоритма проведена процедура оптимизации и определены оптимальные параметры. Далее рассмотрим подробнее каждую из стадий.

5.3.2.1 Сглаживание спектра

По спектру каждого из исследуемых объектов проведено сглаживание с использованием пяти алгоритмов и подбором гиперпараметров. В качестве объекта оценки использовались спектры запрессованных гранул удобрений, как обладающих наибольшей неоднородностью поверхности и химического состава, а, значит, и наибольшей дисперсией данных. Таким образом, именно запрессованные гранулы наиболее требовательны к процедурам предподготовки спектров. Для графического представления полученных метрик использовалось NPK(S) удобрение марки 4-30-15(16), как обладающее наибольшей неоднородностью состава гранул (рисунок 5.22).



а) б)

Рисунок 5.22. Распределение элементов в запрессованных гранулах удобрения, полученное с использованием сканирующего ЭД-РФА микроскопа а) – рентгенограмма б) – наложение распределений элементов, зеленый – фосфор, желтый – калий, синий – сера, оранжевый – кремний.

По полученной рентгенограмме видно, что плотность запрессованных гранул неоднородна. То же можно сказать и про распределение элементов в гранулах.

Значения метрик для всех исследованных объектов одного класса усреднялись и диапазон изменений заносился в соответствующие таблицы. В качестве метрик качества введены и используются такие понятия как:

1. Разница среднего спектра определенного типа удобрения и его среднего сглаженного спектра в импульсах (таблица 5.5).

2. Усредненное относительное среднее квадратичное отклонение (СКО) сглаженных спектров (таблица 5.6).
3. Абсолютное изменение величины интенсивности максимальной характеристической линии (таблица 5.7).
4. Сдвиг пика по оси энергий в относительных процентах от исходного значения (таблица 5.8).

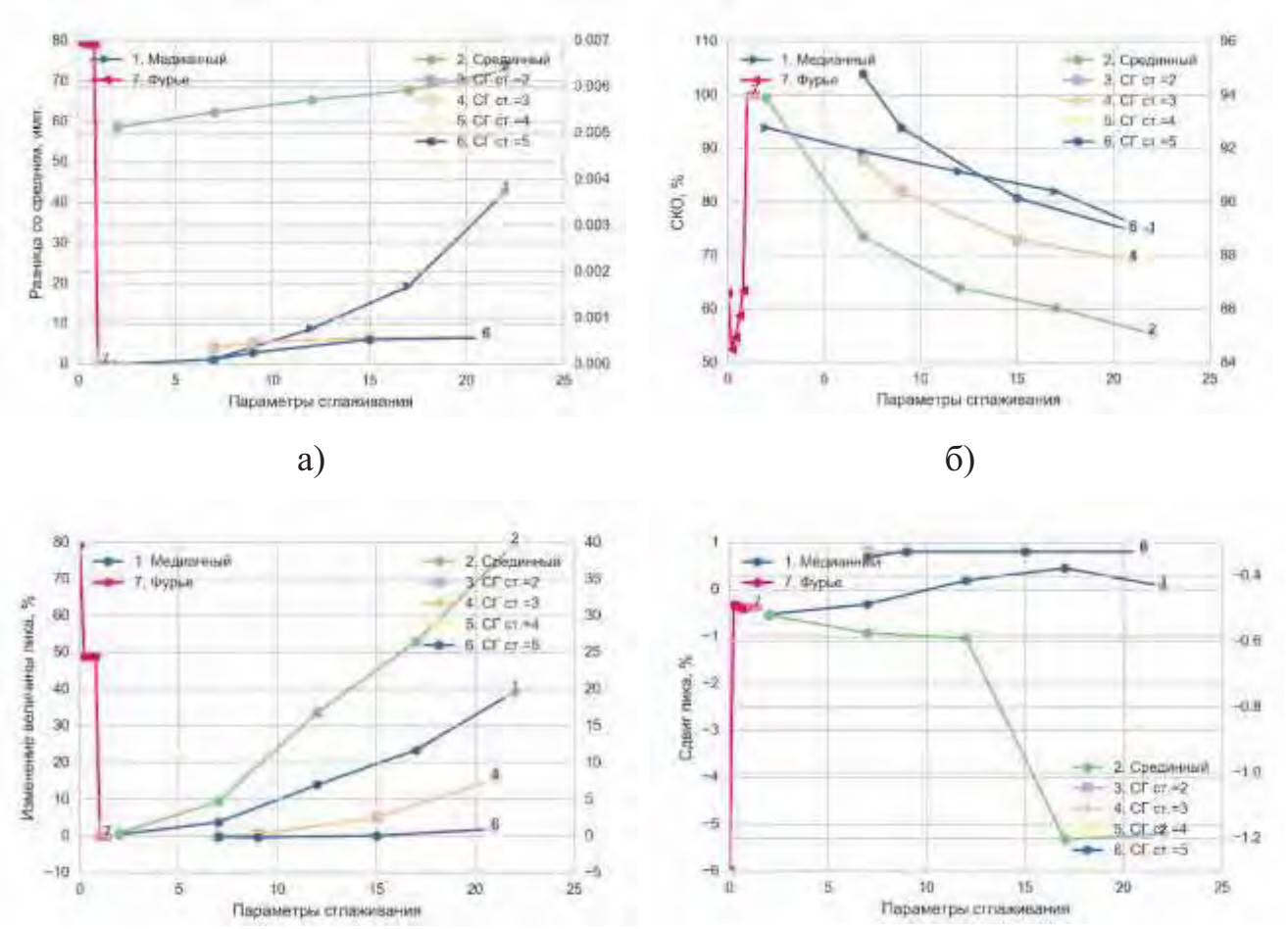


Рисунок 5.23. Оптимизационные кривые различных алгоритмов сглаживания для каждой из метрик.

Таблица 5.5. Значение разницы со средним для каждого типа удобрений, имп.

Удобрение	Медианный	Срединный	Савицкого-Голея	Фурье
NPK(S) 4-20-15(16)	0,005 – 39,7	0,005 – 0,01	0,0001 – 0,0008	$3{,}3\times10^{-13}$ – 76,7
NPK(S) 0-20-20(5)	0,004 – 49,7	0,004 – 0,008	0 – 0,0003	$5{,}4\times10^{-13}$ – 86,5
МАФ 12-52	0,01 – 21,2	0,01 – 0,02	$7{,}2\times10^{-6}$ – 0,0002	0 – 61,06
NPK 15-15-15	0,008 – 23,9	0,006 – 0,015	0 – $3{,}7\times10^{-6}$	0 – 56,9
NP(S) 12-40(10)	0,01 – 18,05	0,01 – 0,03	0 – 0,001	0 – 58,6

По представленным данным наименьшей разницей с исходным спектром обладает алгоритм Савицкого-Голея, что говорит в пользу сохранения большего количества информации. Так же стоит отметить, что наибольший вклад в усредненное значение интенсивности вносят характеристические сигналы в спектре (обычно на несколько порядков превышают средний уровень фона), таким образом данный параметр мало говорит о силе сглаживания фоновой составляющей.

Таблица 5.6. Значение относительного СКО для каждого типа удобрений, %

Удобрение	Медианный	Срединный	Савицкого-Голея	Фурье
NPK(S) 4-20-15(16)	15,0 – 18,0	15,9 – 18,0	16,3 – 18,2	15,0 – 19,6
NPK(S) 0-20-20(5)	8,6 – 12,1	8,8 – 12,1	9,6 – 12,4	6,12 – 14,1
МАФ 12-52	7,5 – 11,1	7,4 – 11,1	8,2 – 11,45	5,4 – 13,6
NPK 15-15-15	18,6 – 23,4	21,2 – 23,4	21,9 – 23,7	13,1 – 25,0
NP(S) 12-40(10)	6,8 – 10,4	7,0 – 10,4	7,9 – 10,7	4,7 – 12,5

Параметр СКО является важной характеристикой говорящей о «количестве» шумовых компонент в исходных данных, включая как характеристическую, так и фоновую составляющие. Наиболее понижающими данный параметр являются алгоритмы Фурье сглаживания, медианного и среднего фильтра. Однако алгоритм Савицкого-Голея со степенью полинома 3 показывает сопоставимые параметры.

Таблица 5.7. Значение абсолютной разницы максимальной интенсивности для каждого типа удобрений, для удобства пересчитанная в %

Удобрение	Медианный	Срединный	Савицкого-Голея	Фурье
NPK(S) 4-20-15(16)	0 – 39,4	0 – 39,9	-0,5 – 7,4	-0,6 – 79,7
NPK(S) 0-20-20	0 – 38,6	0 – 39,1	-0,6 – 7,4	-0,5 – 78,8
МАФ 12-52	0 – 46,4	0 – 44,4	-0,8 – 10,4	-0,8 – 85,8
NPK 15-15-15	0 – 43,6	0 – 42,5	-0,2 – 9,4	-0,8 – 79,8
NP(S) 12-40(10)	0 – 45,8	0 – 44,0	-0,7 – 10,2	-0,9 – 79,6

По данной метрике оценивается изменение формы линии – насколько уменьшается полезный сигнал при проведении сглаживания. Опять наилучшими параметрами обладают алгоритмы Савицкого-Голея. При этом алгоритм Фурье-сглаживания при высоком значении параметра не изменяет величину интенсивности характеристической линии, однако исходя из величины СКО

(максимальная среди всех алгоритмов) можно предположить, что сглаживания, как такового, не происходит.

Таблица 5.8. Значение сдвига максимальной интенсивности по оси энергий для каждого типа удобрений в %

Удобрение	Медианный	Срединный	Савицкого-Голея	Фурье
NPK(S) 4-20-15(16)	-0,21 – 1,05	-0,26 - -0,21	-0,02 - 0	-5,74 – 0
NPK(S) 0-20-20	-2,27 - -0,58	-1,27 - -3,59	-1,11 - -1,11	-1,11 - -6,49
МАФ 12-52	-0,29 – 1,82	-0,43 - -0,29	-0,01 - 0	-0,08 – 0,04
NPK 15-15-15	-0,15 – 1,25	-0,49 – 0,15	-0,03 - 0	-0,06 – 0,96
NP(S) 12-40(10)	-0,17 – 2,06	-0,42 - -0,08	0 - 0	-5,92 - 0

Относительная величина сдвига по оси каналов показывает на сколько алгоритм изменяет положение максимума характеристической линии в процентах. По представленным данным можно сделать вывод, что практически все алгоритмы изменяют положение незначимо (порядка 2% по каналам). Исключение составляют алгоритмы Фурье-преобразования и среднего сглаживания.

Таким образом, оптимальным алгоритмом сглаживания, сохраняющим большую часть полезной информации о спектре неизменной, является алгоритм Савицкого-Голея с полиномом 3 степени и окном в 11 точек.

Представляется интересным проанализировать поведение алгоритма сглаживания на основе Фурье преобразования. Метрики для данного типа сглаживания имеют уникальную структуру и значительно изменяют исходный спектр. Хотя данный тип сглаживание широко используется в современных РФ-спектрометрах (за счет частичного нивелирования базовой линии), но для наших целей он не подходит из-за сильного изменения исходной информации, заложенной в спектре.

5.3.2.2 Алгоритм поиска базовой линии

Для выделения структуры базовой линии и ее последующей аппроксимации полиномом пятой степени используется алгоритм «нулевого фильтра», описанный в пункте 3.2.2 настоящей работы. В качестве сравнения используется алгоритм среднего фильтра, когда все интенсивности каналов меньшие средней,

маркируются как интенсивность базовой линии. Данный алгоритм является наивным и плохо работает со спектрами сложных матриц, однако используется для наглядного сравнения. В качестве метрики качества используется средняя интенсивность базовой линии, усредненная по всем спектрам удобрений одного типа.

В качестве оптимизационных параметров для алгоритма «нулевого фильтра» используются величина окна и константа сдвига полуширины окна относительно рассматриваемого канала. Данные параметры позволяют настроить работу алгоритма для различных профилей линий и типов РФ-спектрометров. Результаты оптимизации данных параметров приведены на рисунке 5.24 (для NPK(S) удобрения марки 4-30-15(16)).

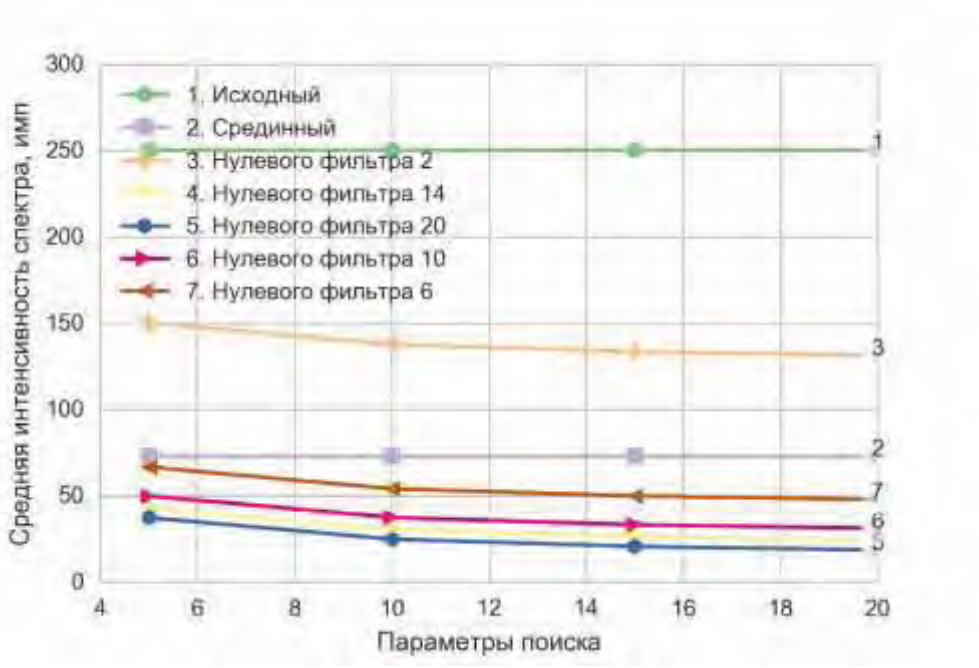


Рисунок 5.24. Оптимизация параметров поиска базовой линии (профиль рассеянного излучения) для NPK(S) удобрений марки 4-30-15(16)

Полученные данные говорят в целом о более эффективной работе алгоритма «нулевого фильтра», чем наивного подхода фильтрации по среднему. Учитывая профиль спектра исследуемых матриц, когда объект содержит много примесных элементов (интенсивности которых меньше основных питательных в 5 – 10 раз), кажется разумным выбор алгоритма с параметрами 10 и 10 для окна поиска и

константы сдвига соответственно, что улучшает качество работы наивного срединного фильтра примерно в 2 раза.

5.3.2.3 Алгоритм выделения характеристических линий

Для выделения характеристической и базовой линий используются алгоритмы дифференцирования с итеративным сглаживанием алгоритмом Савицкого-Голея (параметры 3 и 11) и нулевого фильтра (параметры 10 и 10) для идентификации максимумов найденных характеристических линий. В качестве метрики качества используется усредненное количество найденных пиков:

$$N_p = \sum_N \frac{N_F}{N}$$

где: N_P – метрика качества, N_F – количество пиков, найденное алгоритмом, N – количество исследованных объектов.

Оптимизационные кривые для различных алгоритмов и параметров приведены на рисунке 5.25 и в таблице 5.9.

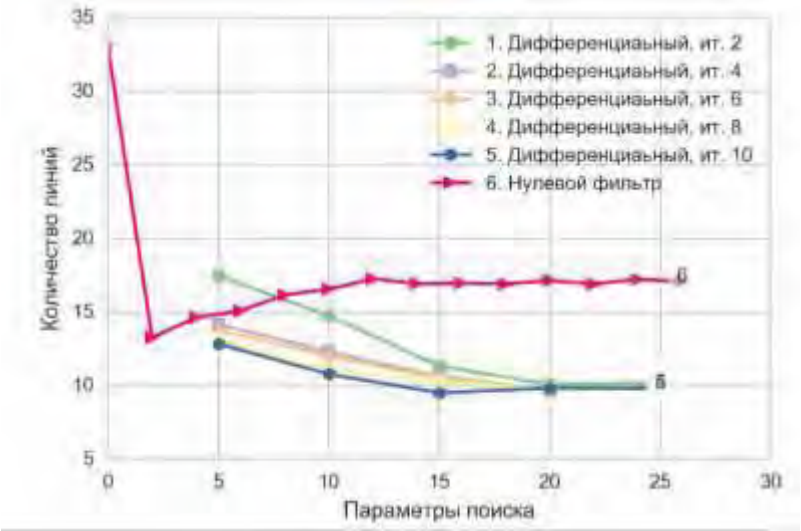


Рисунок 5.25. Оптимизационные кривые для алгоритмов поиска пиков.

Таблица 5.9. Количество найденных линий каждым алгоритмом для выбранного типа удобрений

Удобрение	Дифференциальный	Нулевого фильтра
NPK(S) 4-20-15(16)	9 - 933	13 – 33
NPK(S) 0-20-20(5)	10 - 912	12 - 40
МАФ 12-52	4 - 972	25 - 80
NPK 15-15-15	7 - 952	23 - 58
NP(S) 12-40(10)	3 - 980	24 - 83

Для удобства количество алгоритмов на рисунке ограничено количеством определяемых пиков – оно не должно быть более 70. Завышение работы алгоритма нулевого фильтра (таблица 5.9) вызвано его неустойчивой работой на энергиях спектра, превышающих энергию трубки (в области шумовых данных детектора, которые при дальнейшей работе отсекаются). По представленным данным видно различие в поведении двух типов алгоритмов. Дифференциальный алгоритм ведет себя предсказуемо в зависимости от изменяемых параметров, однако достаточно неустойчив – обладает большим размахом метрики качества. С другой стороны, алгоритм на основе «нулевого фильтра» более устойчив к изменениям параметров и типов проб. На основании введенной метрики выбран алгоритм нулевого фильтра с параметром итераций сглаживания в 6 единиц, как наиболее приближенный к реальному значению количества характеристических линий в спектре.

После нахождения всех характеристических линий и соответствующих им каналов проводиться повторное использование алгоритма нулевого фильтра с загрубленными параметрами для определения границ пиков.

Результат общей работы всех оптимизированных алгоритмов по поиску характеристических и фоновой линий для исследуемых объектов приведен на рисунке 5.26.

Таким образом, используя описанные процедуры становиться возможным провести качественную аппроксимацию базовой линии и определить большую часть характеристических линий спектров, включая когерентное (Релеевское) и не когерентное (Комптоновское) рассеяние излучения анода трубки в полностью автоматическом режиме.

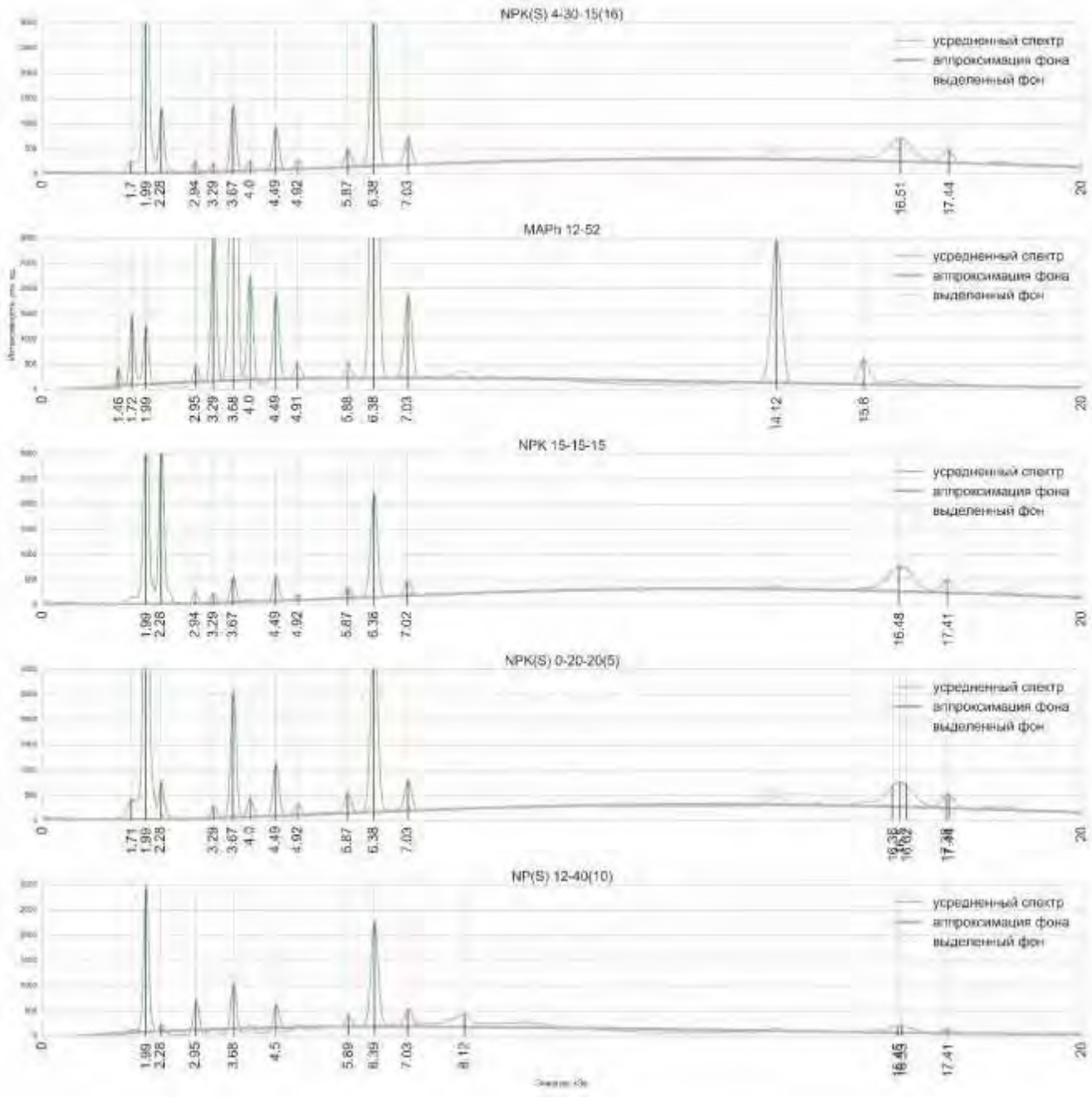


Рисунок 5.26. Работа алгоритма автоматического поиска пиков и аппроксимации базовой линии. Красная прерывистая линия – работа алгоритма нулевого фильтра для выделения фона, синяя утолщенная линия – аппроксимация базовой линии полиномом пятой степени

5.3.2.4 Алгоритм выделения параметров

После определения положения и интенсивности основных характеристических линий в спектре, а также аппроксимации базовой линии полиномом пятой степени, дополнительно рассчитываются значения максимальной интенсивности фона и значение его площади. Положения всех

найденных интенсивностей обобщаются по всем исследованным объектам и, если такой канал с учетом погрешности в 10 % не был учтен в пробе, то он заполняется значением соответствующей интенсивности. Данная процедура позволяет унифицировать аналитическую информацию, поступающую из РФ-спектра по всем исследуемым объектам, что должно положительно сказаться на последующей классификации и регрессии.

Таким образом формируется единый набор признаков для всех объектов. Данные признаки включают в себя:

- обобщённые значения интенсивностей в каналах, определенных оптимизированными алгоритмами;
- значения максимума аппроксимированной интенсивности базовой линии и ее площади.

После заполнения матрицы каждый канал заменяется на соответствующее значение энергии по формуле, приведенной в подразделе 3.3 настоящей работы. Значение энергии заноситься с точностью до 1 знака после запятой. Повторяющиеся значения энергий удаляются, оставшиеся – сравниваются с известными значениями Kα и Lα линий следующих элементов: Si, P, S, Cl, K, Ca, Ti, Fe, Zn, Sr, Mo, Mo не когерентный. При совпадении с заданной точностью значение энергии заменяется на название элемента, при несовпадении признак отбрасывается, что снижает содержание шумовых компонентов в данных.

Далее в матрицу добавляются физические параметры пробы из пункта 5.2 настоящей работы: удельное количество контуров, удельный размер контуров и средняя цветность объекта. Так же вносятся паспортные данные объектов (если известны): значения фракции объекта, тип пробы, марка удобрения.

На следующем этапе производится предобработка данных, включающая в себя нормализацию и кодирование согласно подразделу 3.3 настоящей работы. На выходе получаем матрицу «объекты-признаки» размером 500×50. В данную матрицу занесена основная физико-химическая информация по каждому из

исследованных объектов. Выделенные признаки и их характеристики приведены в таблице 5.10.

Таблица 5.10. Выделенные физико-химические параметры исследованных проб.

Признаки	**Краткое описание**	**Количество в матрице**
Интенсивности характеристических линий основных элементов	Kα и Lα линий следующих элементов: Si, P, S, Cl, K, Ca, Ti, Fe, Zn, Sr, Mo, Mo не когерентный	12
Общая площадь рассеянного излучения	Площадь под аппроксимирующей кривой в диапазоне 0,5 – 25 кэВ.	1
Максимальная интенсивность рассеянного излучения	Максимальная интенсивность аппроксимирующей кривой в диапазоне 0,5 – 25 кэВ.	1
Параметры оптического регистратора	Удельное количество контуров, удельный размер контуров и средняя цветность объекта	3
Паспортные данные объектов (при наличии)	Значения фракции объекта, тип пробы, марка удобрения	7 без использования кодирования

В результате проведенной работы создается база данных исследованных объектов, подготовленная для дальнейшего использования различными алгоритмами для обработки больших данных: классификации, регрессии и т.д.

6 Построение моделей классификации, регрессии и кластеризации

По результатам предыдущей главы получена матрица «объекты-признаки», которая является основой любого метода анализа больших данных. Данную матрицу используют для поиска и определения недостающих данных и построения различных предсказательных моделей. Основные математические алгоритмы построения моделей, использованные в данной работе, описаны в главе 3, а их программная реализация на языке Python 2.7 приведена в приложении А. Стоит отметить, что основное отличие программного подхода от математического заключается в итерациях – именно итеративные приближения заложены в основу практически всех рассматриваемых в работе алгоритмов. В рамках данной главы приводятся результаты поиска, оптимизации и оценки предсказательной силы различных программных подходов. Описанные процедуры проводятся на полученных в результате эксперимента данных, сведенных в единую матрицу «объекты-признаки» (глава 5 настоящей работы).

6.1 Оптический метод

С использованием разработанной оптической установки для исследованных объектов получен набор признаков, которыми являются средняя яркость поверхности объекта в градациях серого, а так же аномалии на карте поверхности гранул и прессованных проб (понятие аномалии и контуры далее используются как идентичные). Поскольку основным направлением работы является построение единого аппаратно-программного комплекса на базе ЭД РФ-спектрометра и оптического регистратора, то в дальнейшем будут рассматриваться только прессованные объекты, если не оговорено обратное.

В качестве признаков объектов выступают три вещественных и два категориальных набора данных. Вещественными признаками являются удельное количество аномалий:

$$N_r = \frac{N}{a \cdot b}$$

где: N_r – удельное количество аномалий, N – количество аномалий, определенных на карте поверхности, a и b – длина и ширина поверхности объекта в пикселях.

И удельная площадь аномалий:

$$S_r = \frac{\sum_{i=1}^{N} S_i}{a \cdot b}$$

где: S_r – удельная площадь аномалий, S_i – площадь аномалии в пикселях, N – количество аномалий, определенных на карте поверхности, a и b – длина и ширина поверхности объекта в пикселях.

А так же средняя яркость поверхности объекта в градациях серого, умноженная на среднюю яркость контрольного белого участка фотографии (левый верхний угол) для нивелирования вариации внешнего освещения:

$$I_g = \frac{\sum_{i=1}^{N} I_i}{N} \times I_w$$

где: I_g – средняя яркость пикселей в градациях серого от 0 до 255, I_i – яркость i-го пикселя в градациях серого от 0 до 255, N – количество пикселей, I_w – средняя яркость белого фона.

Категориальными и, по совместительству, целевыми переменными (для которых проводиться классификация) являются фракция и тип исследуемых объектов. Стоит отметить, что предсказание типа удобрения без знания марки (химического состава по основным питательным элементам) является не информативным. Уже на этом этапе можно сделать вывод, что для классификации удобрений по типу и марке не хватает данных. Дополнительно рассмотрены бинарные параметры, такие как наличие сушки и к.д.. Таким образом, проведя анализ по перечисленным выше параметрам, становиться возможным оценить

значимость каждого из рассчитанных вещественных признаков для классификации.

Каждый объект анализа (таблица 4.1) готовили согласно главе 4 настоящей работы в пяти вариантах: гранулы, порошок фракцией менее 500 мкм и порошок фракцией менее 100 мкм. При этом часть из порошков дополнительно сушили во влагомере в течении 15 минут при температуре 70 ^{0}C. Потеря влаги для каждого объекта останавливалась по достижении 15 минут сушки и не превысила 3 масс. %, что говорит в пользу незначимости данной процедуры для последующего анализа. На рисунке 6.1 приведена карта линейных корреляций рассчитанных признаков для различных исследованных объектов.

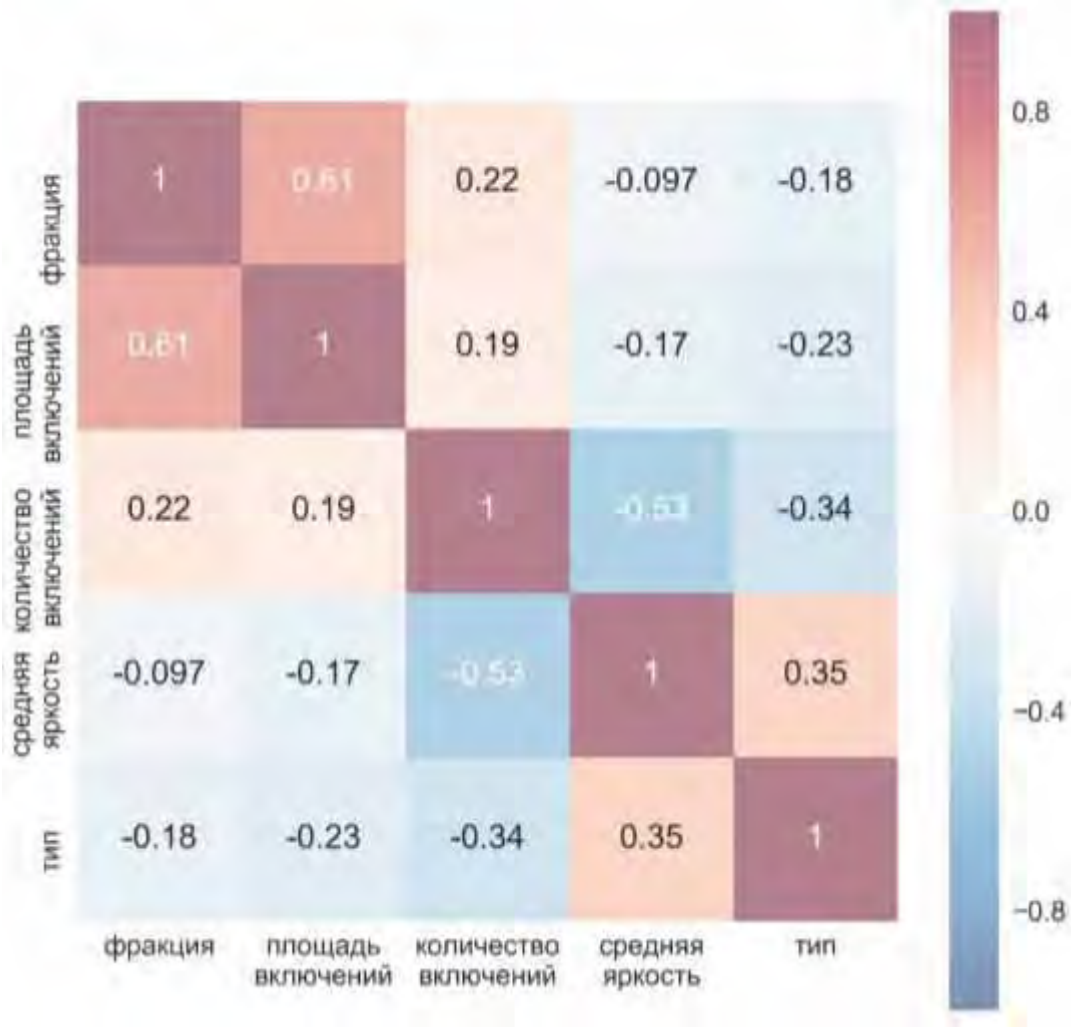
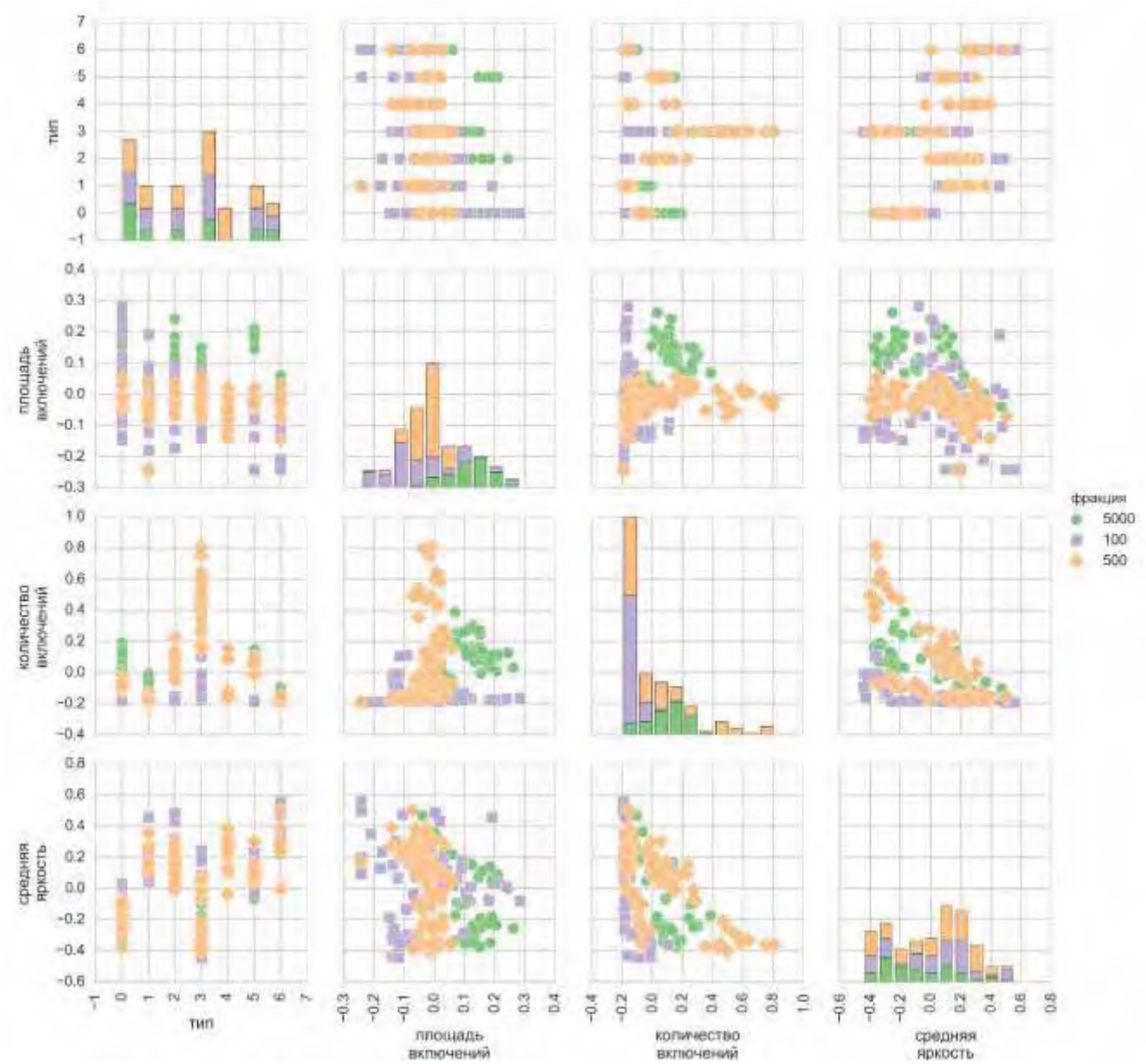
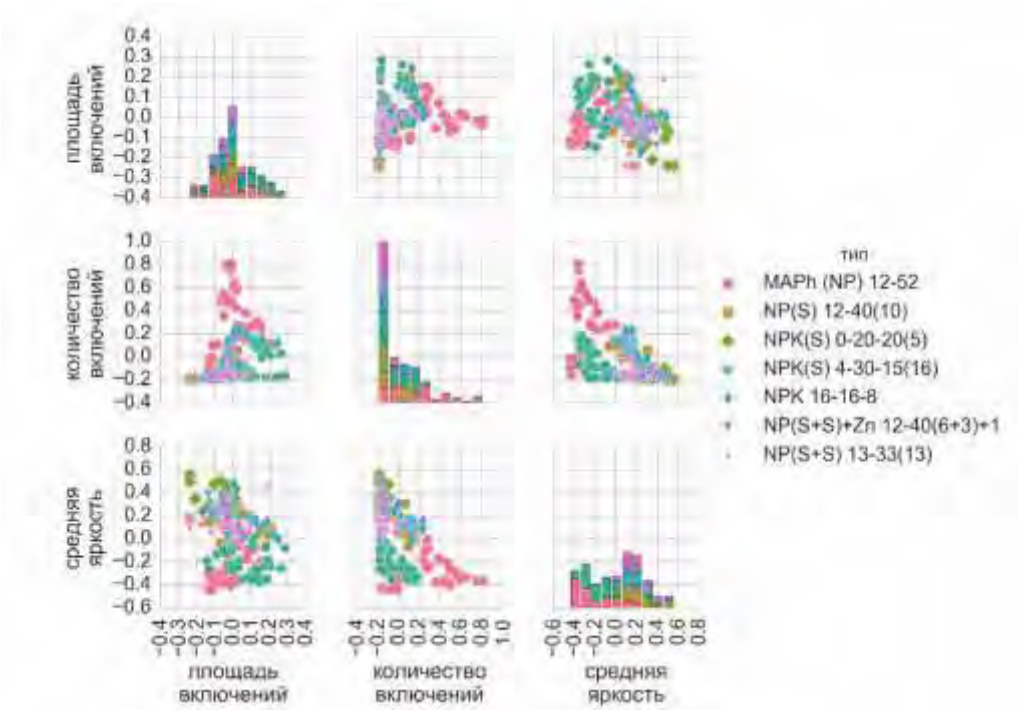


Рисунок 6.1. Карты линейных корреляций по Пирсону для выделенных признаков.

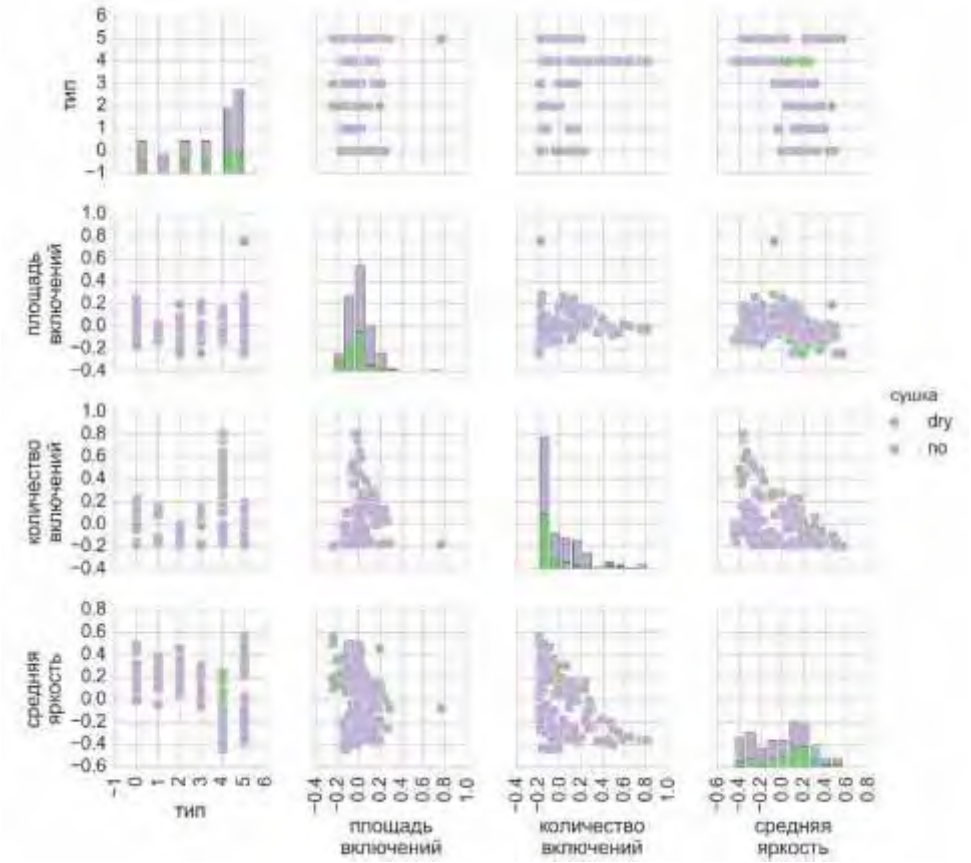
В рамках приведенных данных наблюдается линейная корреляция (61%) между удельной площадью контуров и фракцией объекта. Однако для классификации по типу объекта явной зависимости не наблюдается. Для оценки возможности проведения классификации построены карты бинарных корреляций признаков (рисунок 6.2).



а)



б)



в)

Рисунок 6.2. Карта бинарных корреляций признаков для классификации исследованных объектов а) – по фракции, б) – по типу, в) –наличие предварительной сушки.

По приведенным данным наблюдается определенная взаимосвязь между оптическими параметрами проб и фракцией объекта. В остальных случаях явного разделения целевых параметров не происходит. Важность каждого из признаков для классификации фракции оценена с помощью алгоритма случайного леса (таблица 6.1) для удобрения NPK(S) 4-30-15(16).

Таблица 6.1. Относительная значимость признаков в % для классификации фракции.

Признак	**Значимость, %**
площадь включений	65,70
количество включений	29,39
средняя яркость	4,56

Интересно отметить, что наблюдается полное отсутствие группировки объектов по типу пробы, наличию к.д. и предварительной сушки. Полученные данные согласуются с результатами, представленными в подразделе 5.2 настоящей работы. Подобное поведение может быть обусловлено как не информативностью полученных признаков, так и не значимостью влияния данных целевых переменных на выявленные свойства объектов.

Однако построение корреляций не ограничивается рассмотрением только бинарных признаков и линейных взаимосвязей. Для оценки предсказательной способности выделенных признаков рассмотрим различные подходы к классификации объектов по выделенным целевым переменны. Предварительная обработка данных проведена согласно главе 3 настоящей работы. Использованы нормализация и бинарное кодирование переменных. Для оценки качества используются соответствующие метрики, среди которых точность, полнота и F-мера (F1).

При таком малом количестве признаков неэффективно использовать спрямляющие пространства и понижение размерности данных. Поскольку приведенная далее классификация носит оценочный характер принято решение проводить ее на данных, содержащих максимальное количество исходной информации, поэтому в качестве нормализации использовалась только

нормализация на отрезок. Метрики качества исследуемых алгоритмов с оптимизированными настройками приведены в таблице 6.2.

Таблица 6.2. Сравнение качества предсказания физических параметров пробы

Алгоритм	Показатель	Точность		Полнота		F1	
		Среднее	СКО	Среднее	СКО	Среднее	СКО
Линейная регрессия	предварительная сушка	0,69	0,04	0,63	0,05	0,62	0,05
Линейная регрессия с L1 регуляризацией		0,69	0,04	0,63	0,05	0,62	0,05
Линейная регрессия с L2 регуляризацией		0,69	0,04	0,64	0,05	0,62	0,05
Случайный лес		0,67	0,06	0,64	0,06	0,63	0,06
Наивный Байес		0,44	0,02	0,51	0,02	0,40	0,02
Линейная регрессия	фракция	0,86	0,06	0,86	0,06	0,86	0,06
Линейная регрессия с L1 регуляризацией		0,86	0,06	0,86	0,06	0,86	0,06
Линейная регрессия с L2 регуляризацией		0,84	0,03	0,83	0,05	0,83	0,04
Случайный лес		0,94	0,03	0,93	0,03	0,93	0,03
Наивный Байес		0,66	0,06	0,41	0,06	0,37	0,06
Линейная регрессия	марка удобрения	0,41	0,06	0,54	0,04	0,44	0,04
Линейная регрессия с L1 регуляризацией		0,56	0,08	0,46	0,04	0,44	0,05
Линейная регрессия с L2 регуляризацией		0,54	0,1	0,45	0,05	0,42	0,06
Случайный лес		0,69	0,04	0,68	0,03	0,67	0,03
Наивный Байес		0,22	0,03	0,41	0,03	0,28	0,03

Для сравнения с алгоритмом случайного леса дополнительно рассчитана важность признаков для классификации фракции (как единственно значимой) с использованием линейных алгоритмов (таблица 6.3).

Таблица 6.3. Значимость признаков в долях от общей суммы (100 %)

Линейная регрессия с L2 регуляризацией (Ridge)		**Линейная регрессия с L1 регуляризацией (Lasso)**		**Линейная регрессия**		**Случайный лес**	
Признак	**Значимость, %**	**Признак**	**Значимость, %**	**Признак**	**Значимость, %**	**Признак**	**Значимость, %**
количество контуров	46,31	количество контуров	49,23	количество контуров	41,59	количество контуров	37,37
размер контуров	20,16	размер контуров	5,74	размер контуров	4,96	размер контуров	30,58
средняя яркость	4,31	средняя яркость	0,32	средняя яркость	0,28	средняя яркость	13,89
тип и марка	29,22	тип и марка	44,71	тип и марка	53,17	тип и марка	18,19

Для линейных алгоритмов значимость признаков рассчитывалась по формуле:

$$im_j = \frac{w_j}{\sum_1^n w_i} \times 100$$

где: im_j – «важность» j-ого признака в %, w_j – вес j-ого признака в уравнении регрессии, w_i – вес i-ого признака в уравнении регрессии.

Поправка на дисперсию признаков для оценки значимости не вводилась, поскольку данные нормализованы.

В результате классификации не установлено значимой связи между физическими признаками, полученными в ходе проведения оптического анализа, типа и наличия предварительной сушки объектов. С другой стороны, достигнута точность классификации фракцией объекта в 94 %, что является хорошим результатом. Наилучший результат показал алгоритм случайного леса с параметрами: количество «деревьев» - 82, максимальное количество признаков – не ограничено, максимальная глубина каждого дерева – не ограничено, бутстрап – нет, взвешенные кассы – да. Наилучшим линейным алгоритмом является линейная классификация с L1 регуляризацией или без нее с точностью в 86 %. Данный результат в какой-то мере предсказуем, поскольку по предварительной оценке бинарных корреляций не было выявлено линейно разделимых классов. Наихудший результат показал наивный Байесовский классификатор, что может быть вызвано

как малым размером выборки (меньше 1000 объектов), так и несоблюдением теоремы Байеса для полученных данных. Интересно отметить, что наиболее значимым признаком является среднее удельное количество контуров, что противоречит первоначальному предположению большей информативности размера контуров перед их количеством. Скорее всего данное явление обусловлено большей универсальностью данного параметра при работе с различными марками удобрений.

Взаимосвязи с наличием предварительной сушки, типа и марки объектов установлены на уровне не превышающим 70 %, что говорит о недостаточной информативности выбранных признаков или незначимости влияния данных параметров на морфологию поверхности излучателя.

По результатам работы оптического регистратора становиться очевидным, что для эффективного контроля качества производимых минеральных удобрений необходим дополнительный источник информации.

6.2 Спектральные признаки

Таким источником дополнительной информации о химическом составе пробы является энергодисперсионный РФ спектрометр. По признакам, выделенным из РФ-спектра так же имеет смысл провести бинарную классификацию для оценки влияния пробоподготовки и типа объекта на характеристические линии основных питательных элементов.

По исследованным объектам построена карта бинарных классификаций между маркой удобрения и аналитическими сигналами основных питательных элементов (рисунок 6.3).

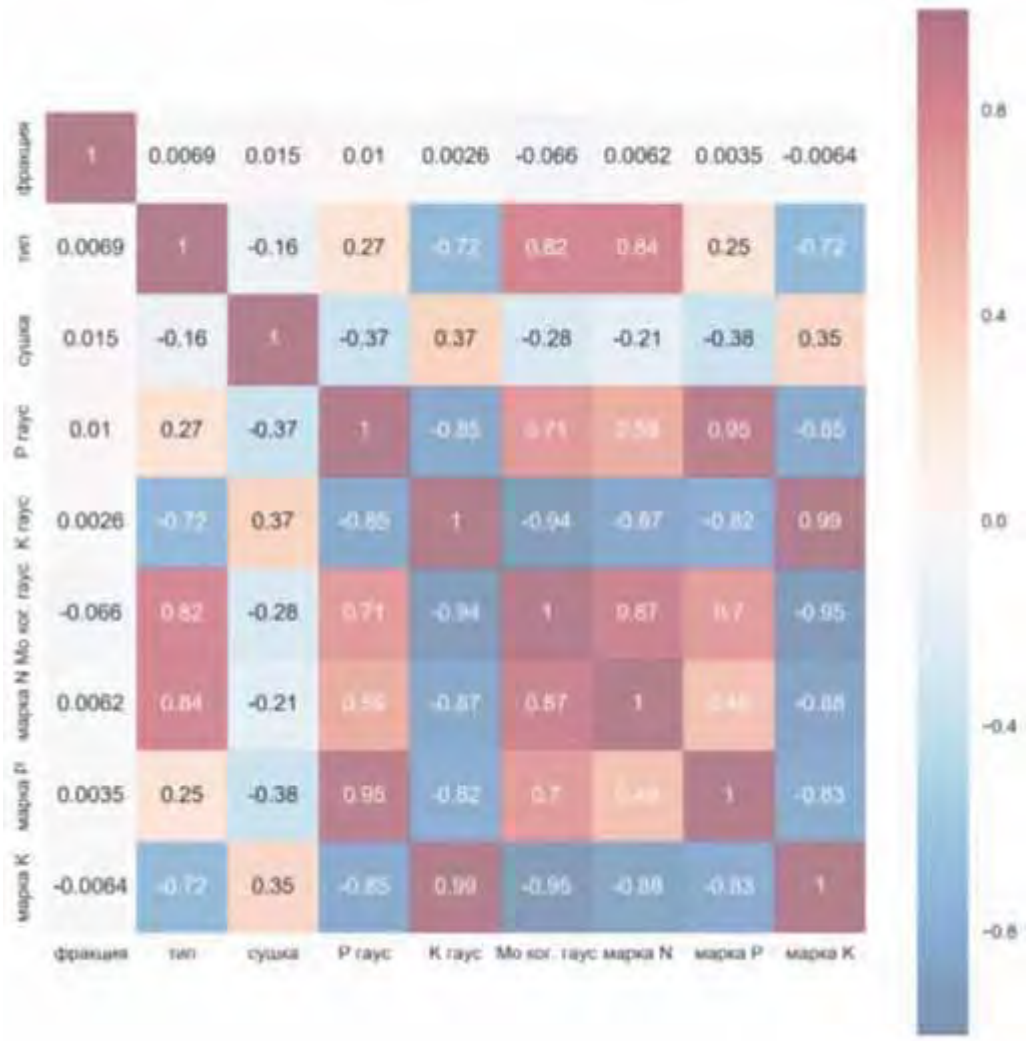


Рисунок 6.3. Бинарные классификации исследованных объектов.

Характеристические линии элементов аппроксимировались гауссианами и вычислялась соответствующая интенсивность. По представленным данным наблюдается множество сильных корреляций, что говорит о большом количестве линейных связей между полученными признаками. Данный факт не очень хорош для статистического анализа, однако не противоречит физическому смыслу. Действительно, при получении ЭД РФ спектра наблюдается сильная связь аналитических сигналов с фундаментальными параметрами и, как следствие, друг с другом. Так же многие свойства спектров зависят от типа удобрения и его марки. Данный факт говорит в пользу использования регуляризации при проведении дальнейшего статистического анализа. Интересным показателем является

отсутствие сильной связи признаков с фракцией объекта, что говорит в пользу совместного использования спектрального и оптического контроля.

Для параметра наличия предварительной сушки, так же как и в случае оптического регистратора, не выявлено значимо изменяющихся признаков, что говорит в пользу отсутствия влияния данного параметра на результаты РФ измерений. Для подтверждения установленного факта проведена статистическая оценка среднего, доверительного интервала для среднего и размаха для самого неоднородного удобрения NPK(S) 4-30-15(16) (таблицы 6.4 и 6.5).

Дополнительно проведено статистическое сравнение различия в средних интенсивностях характеристических линий основных питательных элементов от наличия предварительной сушки. Исследованные данные проверены на нормальность распределения с помощью квантиль-квантиль графиков для всех объектов (рисунок 6.4) и в рамках конкретного удобрения фракции 500 мкм (рисунок 6.5).

Таблица 6.4. Статистические показатели не высушенных объектов

Интенсивность пика	гранулы NPK(S)			NPK(S) < 500 мкм			NPK(S) < 100 мкм		
	среднее, имп.	СКО, %	размах, %	среднее, имп.	размах, %	СКО, %	среднее, имп.	размах, %	СКО, %
максимум фона	198,7	5,120	15,636	203,9	1,121	4,505	204,2	2,137	7,696
фосфора	5316,7	6,128	21,316	5002,1	1,336	5,611	5021,9	2,141	7,832
серы	2506,0	7,324	35,408	2672,2	1,859	7,347	2688,5	2,976	9,696
калия	12827,4	7,098	32,519	14438,9	1,345	6,085	14406,7	2,740	9,692
железа	1092,1	7,623	27,286	1195,8	2,329	7,582	1176,6	3,967	15,411
некогерентного рассеяния	307,9	6,814	23,705	317,9	3,346	12,373	321,6	4,737	19,592
когерентного рассеяния	297,5	6,395	22,405	304,1	2,980	11,070	305,2	5,030	19,330

Таблица 6.5. Статистические показатели высушенных объектов

Интенсивность пика	NPK(S) < 500 мкм			NPK(S) < 100 мкм		
	среднее, имп.	размах, %	СКО, %	среднее, имп.	размах, %	СКО, %
максимум фона	204,5	0,285	1,046	206,7	0,677	2,733
фосфора	5034,6	0,886	3,330	5046,8	0,853	2,721
серы	2674,6	0,973	3,377	2706,8	1,009	3,916
калия	14510,4	0,752	2,481	14550,4	0,543	2,098
железа	1229,6	6,261	22,257	1196,3	2,754	10,115
некогерентного рассеяния	320,5	3,191	12,270	323,3	2,901	9,896
когерентного рассеяния	295,1	3,328	13,442	316,4	14,54	4,330

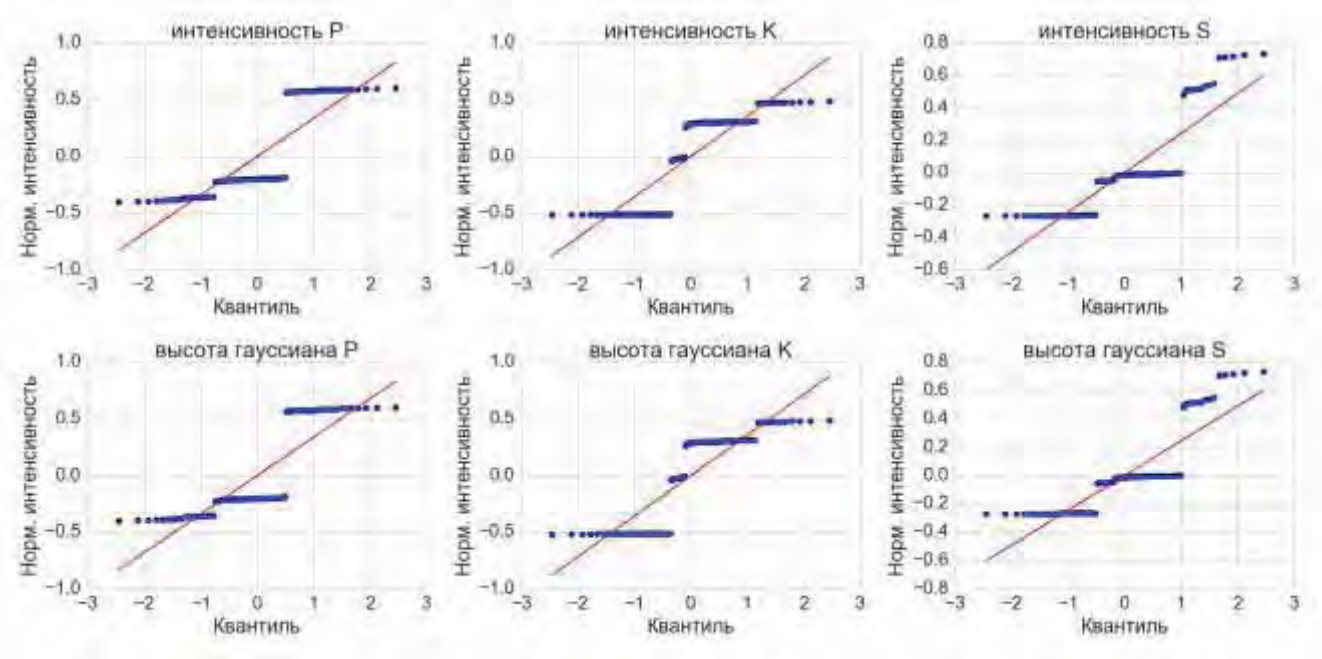


Рисунок 6.4. Квантиль-квантиль графики для нормализованных интенсивностей характеристических линий P, S и K рассчитанных как средние по линии и при аппроксимации линии по Гауссу для всех объектов.

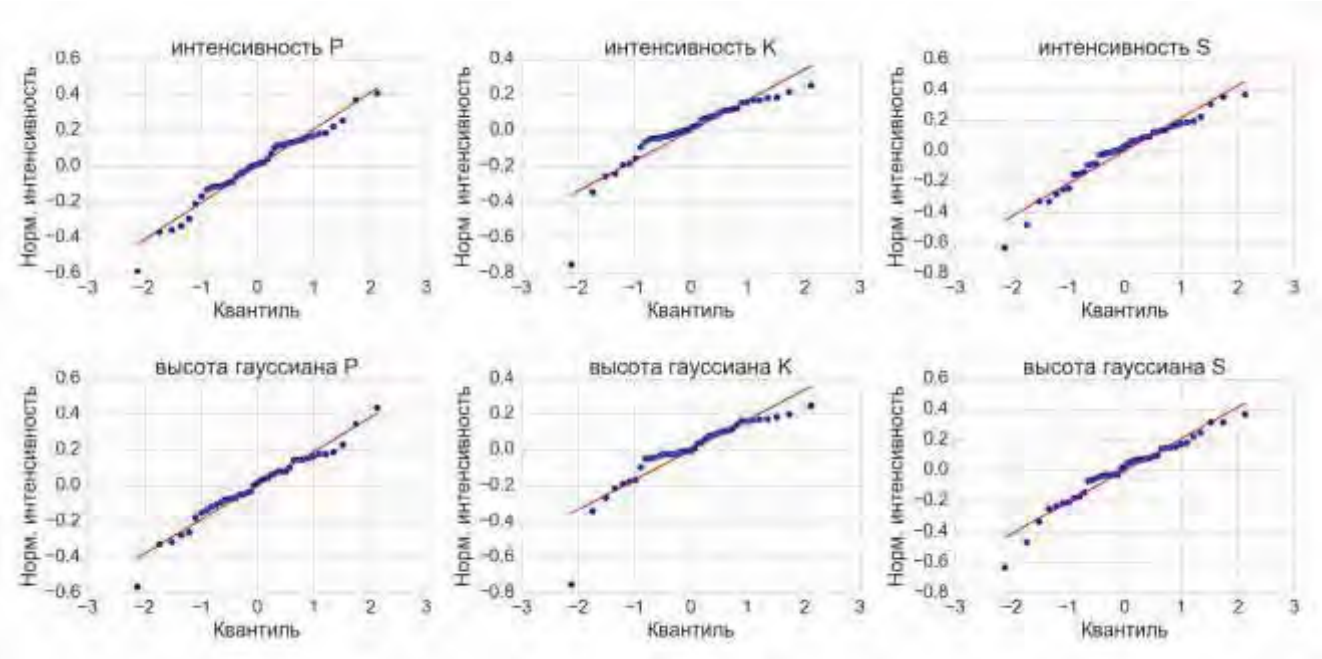


Рисунок 6.5. Квантиль-квантиль графики для нормализованных интенсивностей характеристических линий P, S и K рассчитанных как средние по линии и при аппроксимации линии по Гауссу для NPK(S) удобрения марки 4-30-15(16).

Можно заметить, что в рамках конкретной марки удобрения распределение близко к нормальному, однако наблюдаются некоторые выбросы. В то время как для обобщенного случая распределение не является нормальным. Данный

результат достаточно закономерен, поскольку каждое удобрение обладает своим, зачастую уникальным, составом и производится по своей технологической схеме. Так же можно обратить внимание, что расчет интенсивностей с аппроксимацией характеристических линий по Гауссу дает распределение более приближенное к нормальному. Далее в работе будет использоваться интенсивность линий, рассчитанная по аппроксимации по Гауссу, если не оговорено обратное. Таким образом в качестве общего статистического критерия оценки различия средних используется критерий Манна-Уитни (таблица 6.6), поскольку он не требует нормального распределения в данных.

Таблица 6.6. Статистический анализ влияния наличия предварительной сушки на результаты РФ-анализа основных питательных элементов.

Фракция	**p-уровень значимости**		
	интенсивность фосфора	**интенсивность калия**	**интенсивность серы**
100	0,104	0,261	0,288
500	0,366	0,272	0,458

Исходя из статистических тестов обнаружены статистически значимые различия (для 95 % p – уровня значимости в 0,05). Однако интенсивности сигналов достаточно высоки и подобное отличие не должно сказаться на общем качестве РФ-анализа. Так же 95 % доверительный интервал для среднего, в случае с объектами с предварительной сушкой, всегда пересекается с доверительным интервалом объектов без предварительной сушки и оказывается большим по величине (таблица 6.7), что говорит о большем разбросе данных после проведения сушки. Полученные аномальные результаты могут быть обусловлены более плохой воспроизводимостью анализа за счет интенсивного поглощения влаги из окружающей среды поверхностью объектов и более низким качеством прессования поверхности таблетки.

Таблица 6.7. 95 % доверительный интервал для среднего объектов с предварительной сушкой и без.

Фракция, мкм	**Сушка**	**P**	**S**	**K**	**Разница в размере диапазона «с сушкой – без сушки», %**		
					P	**S**	**K**
100	нет	4269,98 – 4600,44	2589,22 – 3504,16	12740,27 – 14408,32	43,53	52,86	58,81
	да	4028,08 – 4502,40	2646,60 – 4045,21	12107,28 – 14756,33			
500	нет	4238,21 – 4583,57	2562,73 – 3416,35	12999,76 – 14,522,04	32,48	38,23	41,38
	да	4047,63 – 4505,16	2594,88 – 3774,87	12659,92 – 14812,17			

Интересно отметить, что разница между диапазонами и средними значениями для фракции 500 мкм меньше чем для 100 мкм. Полученные данные свидетельствует в пользу отказа от проведения предварительной сушки объектов.

В то же время, с использованием выделенных признаков, становиться возможно предсказать тип и марку удобрения. Для этого использованы и оптимизированы те же алгоритмы классификации, что приведены в пункте 6.1 настоящей работы. Аналогично, классификация проводилась только для данных, нормализованных на отрезок. Результаты работы оптимизированных алгоритмов занесены в таблицу 6.8. Оценки проводились по стратегии кросс-валидации на отложенном тесте (30 % от выборки) по случайным подвыборкам с сохранением распределения классов 10 раз, результат усреднялся.

Таблица 6.8. Сравнение качества предсказания физических параметров пробы

Алгоритм	Показатель	Точность		Полнота		F1	
		Среднее	СКО	Среднее	СКО	Среднее	СКО
Линейная регрессия	Фракция	0,81	0,04	0,80	0,04	0,80	0,04
Линейная регрессия с L1 регуляризацией		0,80	0,04	0,80	0,04	0,80	0,04
Линейная регрессия с L2 регуляризацией		0,79	0,04	0,78	0,04	0,78	0,04
Случайный лес		0,93	0,03	0,93	0,03	0,93	0,03
Наивный Байес		0,19	0,02	0,29	0,02	0,21	0,02
Линейная регрессия	N	0,9980	0,004	0,9950	0,02	0,9960	0,01
Линейная регрессия с L1 регуляризацией		0,9986	0,004	0,9950	0,02	0,9965	0,01
Линейная регрессия с L2 регуляризацией		0,9986	0,004	0,9950	0,02	0,9965	0,01
Случайный лес		0,9992	0,002	0,9968	0,009	0,9979	0,006
Наивный Байес		0,7132	0,02	0,7470	0,02	0,6747	0,02
Линейная регрессия	P	0,9988	0,003	0,9960	0,01	0,9972	0,008
Линейная регрессия с L1 регуляризацией		0,9983	0,004	0,9940	0,01	0,9983	0,009
Линейная регрессия с L2 регуляризацией		0,9988	0,003	0,9960	0,01	0,9972	0,008
Случайный лес		1,00	0	1,00	0	1,00	0
Наивный Байес		0,7785	0,03	0,7544	0,03	0,7735	0,03
Линейная регрессия	K	0,9985	0,004	0,9937	0,02	0,9957	0,01
Линейная регрессия с L1 регуляризацией		1,00	0	1,00	0	1,00	0
Линейная регрессия с L2 регуляризацией		1,00	0	1,00	0	1,00	0
Случайный лес		1,00	0	1,00	0	1,00	0
Наивный Байес		0,7954	0,02	0,7833	0,02	0,7800	0,02
Линейная регрессия	S	0,9905	0,01	0,9921	0,01	0,9908	0,01
Линейная регрессия с L1 регуляризацией		0,9988	0,002	0,9950	0,02	0,9969	0,009
Линейная регрессия с L2 регуляризацией		1,00	0	1,00	0	1,00	0
Случайный лес		0,9928	0,01	0,9700	0,05	0,9782	0,04
Наивный Байес		0,7000	0,05	0,6454	0,05	0,6453	0,05

Дополнительно рассчитана важность признаков для линейных алгоритмов аналогично пункту 6.1 настоящей работы (таблица 6.9).

Таблица 6.9. Значимость признаков в долях от общей сумма (100 %)

	Линейная регрессия с L2 регуляризацией (Ridge)		Линейная регрессия с L1 регуляризацией (Lasso)		Логистическая регрессия		Случайный лес	
	Признак	**Значимость, %**	**Признак**	**Значимость, %**	**Признак**	**Значимость, %**	**Признак**	**Значимость, %**
Фракция	K	13,95	K	11,82	K	12,88	P	10,35
	Fe	10,27	Fe	10,78	Fe	8,59	Ca	10,09
	площадь фона	8,09	площадь фона	9,80	площадь фона	8,57	Fe	9,17
	Cl	8,02	Cl	7,72	Cl	8,14	K	9,15
	Тип	26,05	Тип	26,40	Тип	27,33	Cl	8,60
N	Sr	14,71	Si	18,05	Si	14,63	Cl	13,85
	Ca	12,32	Ca	14,71	Ca	14,20	K	13,04
	Cl	12,01	Sr	14,21	Sr	12,51	Sr	12,39
	площадь фона	7,31	Cl	10,75	Cl	9,91	площадь фона	11,08
	P	7,00	K	7,18	K	6,64	Mo когерентный	9,62
P	Ca	12,52	Si	11,03	Si	13,97	Ca	20,58
	Sr	11,62	Cl	10,64	Ca	12,50	Mo	10,38
	Cl	10,48	Sr	10,43	Sr	11,18	K	10,12
	K	7,42	сушка	6,01	Cl	10,00	площадь фона	10,08
	P	6,67	Ca	5,57	Ti	6,72	Cl	9,24
K	Si	18,64	Si	18,64	Si	15,31	K	19,70
	Sr	14,40	Sr	14,40	Sr	12,13	Cl	15,35
	Ca	13,48	Ca	13,48	Cl	9,97	Ca	13,35
	Cl	11,18	Cl	11,18	Ca	7,52	Fe	12,02
	K	6,61	K	6,61	K	5,86	площадь фона	7,93
S	Sr	13,94	Sr	17,19	Si	14,43	Mo	29,09
	Ca	12,45	Cl	16,48	Sr	13,80	Mo когерентный	28,92
	Cl	12,14	Si	12,94	Cl	12,46	Ca	20,91
	площадь фона	7,37	Ca	10,84	Ca	10,94	Fe	15,98
	K	7,19	Ti	7,95	Ti	7,70	Mn	4,38

Таким образом становиться возможным точное определение марки и типа исследуемого удобрения. Полученные данные говорят о незначимости фракции и предварительной сушки для классификации при наличии большого количества

данных. Однако при решении аналитических задач для удобрений конкретной марки или задач повышенной точности (регрессия), особенно в условиях недостаточного количества исходных данных, значение фракции необходимо, для чего и будет использован оптический способ контроля. Так же выявлено влияние основных питательных элементов на фракцию, что может свидетельствовать и об обратной взаимосвязи.

6.3 Объединенный набор данных

Исходя из предыдущих пунктов настоящей главы, становиться очевидна недостаточная информативность отдельно оптического и спектрального методов для предсказания совокупности физических и химических свойств исследуемых объектов. В настоящем разделе продолжается описание исследования возможности определения фракции объекта, типа и марки удобрений по основным питательным элементам и их химическому составу. Для этого объединены данные оптической и спектральной установок, проведен подбор алгоритмов классификации и регрессии. Дополнительно рассмотрена работа алгоритмов при двух различных схемах нормализации данных (на отрезок и Z-преобразование согласно главе 3 настоящей работы). Показатели качества предсказаний на объединенном наборе данных приведены в таблице 6.10.

Интересно отметить, что линейная регрессия без регуляризации и нелинейный алгоритм случайного леса лучше работают с данными, нормализованными на отрезок в то время как линейная регрессия с регуляризацией – с данными нормированными на среднее и дисперсию (z-преобразование).

Таблица 6.10. Сравнение качества предсказания физических параметров пробы с нормализацией данных на отрезок

Алгоритм	Показатель	Нормализация на отрезок						Z-преобразование					
		Точность		Полнота		F1		Точность		Полнота		F1	
		Среднее	СКО	Среднее	СКО	Среднее	СКО	Среднее	СКО	Среднее	СКО	Среднее	СКО
Линейная регрессия	Фракция	0,9268	0,03	0,9212	0,03	0,9240	0,03	0,9194	0,03	0,9142	0,04	0,9151	0,04
Линейная регрессия с L1 регуляризацией (Lasso)		0,9317	0,03	0,9223	0,03	0,9251	0,03	0,9335	0,03	0,9249	0,03	0,9277	0,03
Линейная регрессия с L2 регуляризацией (Ridge)		0,9187	0,02	0,9109	0,02	0,9133	0,02	0,9070	0,03	0,8968	0,04	0,8990	0,03
Случайный лес		0,9835	0,02	0,9855	0,02	0,9840	0,02	0,9767	0,02	0,9785	0,02	0,9772	0,02
Наивный Байес		0,5081	0,06	0,5778	0,06	0,5332	0,06	0,6768	0,04	0,6518	0,05	0,6389	0,05
Линейная регрессия	N	0,9971	0,005	0,9900	0,01	0,9931	0,01	0,9965	0,008	0,9967	0,008	0,9964	0,007
Линейная регрессия с L1 регуляризацией		0,9986	0,004	0,9950	0,02	0,9965	0,01	0,9944	0,01	0,9984	0,003	0,9963	0,007
Линейная регрессия с L2 регуляризацией		0,9986	0,004	0,9950	0,02	0,9965	0,01	0,9993	0,002	0,9975	0,007	0,9983	0,005
Случайный лес		1,0	0	1,0	0	1,0	0	0,9929	0,01	0,9922	0,01	0,9922	0,009
Наивный Байес		0,6496	0,02	0,7963	0,03	0,7112	0,03	0,9505	0,02	0,9444	0,02	0,9419	0,02
Линейная регрессия	P	0,9978	0,007	0,9980	0,006	0,9978	0,007	0,9978	0,007	0,9980	0,006	0,9978	0,007
Линейная регрессия с L1 регуляризацией		0,9978	0,007	0,9980	0,006	0,9978	0,007	0,9978	0,007	0,9980	0,006	0,9978	0,007
Линейная регрессия с L2 регуляризацией		0,9978	0,007	0,9980	0,006	0,9978	0,007	1,0	0	1,0	0	1,0	0
Случайный лес		1,0	0	1,0	0	1,0	0	0,9933	0,01	0,9963	0,008	0,9944	0,01
Наивный Байес		0,7603	0,01	0,7037	0,02	0,6071	0,02	0,9808	0,01	0,9709	0,02	0,9800	0,02

Алгоритм	Показатель	Нормализация на отрезок						Z-преобразование					
		Точность		Полнота		F1		Точность		Полнота		F1	
		Среднее	СКО	Среднее	СКО	Среднее	СКО	Среднее	СКО	Среднее	СКО	Среднее	СКО
Линейная регрессия	K	0,9985	0,003	0,9937	0,01	0,9959	0,008	1,0	0	1,0	0	1,0	0
Линейная регрессия с L1 регуляризацией		0,9985	0,004	0,9937	0,02	0,9957	0,01	0,9985	0,004	0,9938	0,02	0,9957	0,01
Линейная регрессия с L2 регуляризацией		1,0	0	1,0	0	1,0	0	1,0	0	1,0	0	1,0	0
Случайный лес		1,0	0	1,0	0	1,0	0	1,0	0	1,0	0	1,0	0
Наивный Байес		0,7134	0,005	0,7593	0,008	0,6684	0,007	0,9656	0,02	0,9630	0,03	0,9613	0,03
Линейная регрессия	S	0,9942	0,02	0,9978	0,006	0,9956	0,01	0,9852	0,02	0,9914	0,02	0,9872	0,02
Линейная регрессия с L1 регуляризацией		0,9848	0,03	0,9957	0,008	0,9887	0,02	0,9942	0,02	0,9978	0,006	0,9956	0,01
Линейная регрессия с L2 регуляризацией		0,9979	0,004	0,9850	0,03	0,9899	0,02	0,9859	0,02	0,9964	0,006	0,9903	0,02
Случайный лес		0,9979	0,004	0,9850	0,03	0,9899	0,02	0,9967	0,005	0,9750	0,04	0,9830	0,03
Наивный Байес		0,8193	0,003	0,8148	0,004	0,7372	0,004	0,9846	0,006	0,9805	0,009	0,9834	0,009
Линейная регрессия	Сушка	0,7384	0,07	0,7281	0,06	0,7294	0,06	0,7366	0,07	0,7303	0,07	0,7237	0,05
Линейная регрессия с L1 регуляризацией		0,7386	0,07	0,7307	0,06	0,7308	0,06	0,7333	0,06	0,7414	0,05	0,7344	0,06
Линейная регрессия с L2 регуляризацией		0,7386	0,08	0,6627	0,04	0,6846	0,05	0,6913	0,09	0,7117	0,1	0,6869	0,1
Случайный лес		0,7760	0,05	0,7793	0,06	0,7737	0,05	0,7919	0,07	0,7383	0,08	0,7550	0,08
Наивный Байес		0,8171	0,003	0,7407	0,008	0,7304	0,007	0,7071	0,008	0,5021	0,01	0,5502	0,01

Как правило, наилучший результат по F-мере дает нелинейный алгоритм случайного леса с данными, нормализованными на отрезок, что говорит о наличии определенных нелинейных закономерностях в данных. Так же данное явление свидетельствует в пользу информативности разброса данных для их классификации. Тем не менее в некоторых случаях (P и K) качество работы регрессии с регуляризацией на z-преобразованных данных не сильно уступает алгоритму случайного леса, а для N и S даже превосходит его. В целом классификация работает с точностью не менее 98 % даже для фракции, что удовлетворяет требованиям технологического контроля. Исключение составляет параметр наличия предварительной сушки по причинам, рассмотренным ранее в настоящей работе.

Далее приведена оценка значимости признаков для классификации, рассчитанная по значению весов регрессионного уравнения, аналогично пункту 6.1 настоящей работы для двух типов нормализации данных (таблица 6.11). Стоит отметить, что использование различной технологии нормализации данных приводит к изменению значимости тех или иных факторов для классификации. Однако данное явление достаточно закономерно.

Таблица 6.11. Значимость признаков для классификации, рассчитанная по линейным алгоритмам регрессии.

	Нормализация на отрезок								Z-преобразование							
	Линейная регрессия		Линейная регрессия с L1 регуляризацией		Линейная регрессия с L2 регуляризацией		Случайный лес		Линейная регрессия		Линейная регрессия с L1 регуляризацией		Линейная регрессия с L2 регуляризацией		Случайный лес	
	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %
Фракция	количество контуров	18,18	количество контуров	24,72	K	15,22	размер контуров	25,70	количество контуров	20,72	количество контуров	25,72	K	20,84	количество контуров	26,81
	K	14,58	K	12,84	количество контуров	14,17	количество контуров	21,87	K	16,23	K	15,87	количество контуров	11,03	размер контуров	25,25
	P	10,09	S	10,07	P	10,32	K	6,17	Mo	8,86	Cl	9,08	Cl	9,13	Ca	5,78
	площадь фона	8,77	Mo	8,75	площадь фона	8,93	Ca	5,78	S	7,65	Mo	8,77	P	8,18	K	5,41
	Mo	7,71	P	8,55	S	7,98	Cl	5,03	P	7,51	S	7,66	Fe	7,95	Fe	5,38
N	средняя яркость	10,25	средняя	15,12	Sr	11,32	Cl	15,56	Si	13,41	Si	16,19	K	22,38	площадь фона	16,67
	Cl	8,87	Sr	14,10	Ca	10,80	K	13,90	Ca	13,28	Ca	15,26	Ca	14,13	Cl	16,67
	Sr	7,56	Cl	12,93	Cl	9,96	Sr	13,77	Sr	10,24	Cl	10,41	Sr	12,61	Fe	16,67
	P	6,85	Ca	9,26	средняя яркость	8,73	Ca	10,65	Cl	9,38	Sr	10,36	Cl	9,63	Mo	16,67
	Si	6,66	Si	8,10	Ca	17,69	Mo	9,09	K	8,90	K	9,55	P	7,77	Sr	16,67

	Нормализация на отрезок								Z-преобразование							
	Линейная регрессия		Линейная регрессия с L1 регуляризацией		Линейная регрессия с L2 регуляризацией		Случайный лес		Линейная регрессия		Линейная регрессия с L1 регуляризацией		Линейная регрессия с L2 регуляризацией		Случайный лес	
	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %
P	Ca	15,13	Ca	17,25	K	15,42	Ca	17,04	Si	11,95	Ca	13,28	K	16,02	Ca	19,05
	Si	13,36	K	13,63	Sr	13,34	Si	14,77	Ca	11,83	Si	13,05	Ca	13,85	Si	14,68
	K	11,52	Si	13,15	Si	12,41	Cl	10,70	K	11,74	K	12,48	Sr	11,50	Mo когерентный	12,31
	Sr	9,99	Sr	11,36	Cl	11,20	K	10,43	Sr	10,47	Sr	10,47	Si	9,42	площадь фона	8,31
	Cl	9,20	Cl	9,08	площадь фона	5,38	площадь фона	10,23	Cl	7,50	S	7,59	Cl	7,11	Mo	7,94
K	средняя яркость	10,58	средняя яркость	14,13	Sr	11,24	Cl	14,31	Ca	14,26	средняя яркость	14,44	Ca	18,50	Cl	14,11
	Sr	8,76	Sr	13,47	Ca	11,23	K	12,29	Si	13,52	Sr	13,05	K	14,87	K	12,49
	Ca	8,23	Ca	10,62	Cl	9,94	Sr	12,15	K	10,20	Si	9,75	Sr	11,48	Fe	11,94
	Cl	8,03	Cl	9,88	средняя яркость	8,51	Ca	12,14	Cl	9,97	Ca	8,92	Si	11,44	Mo	10,66
	площадь фона	7,05	площадь фона	8,45	площадь фона	7,45	Mo	10,61	Sr	9,74	Cl	7,56	Cl	9,40	Sr	10,62

	Нормализация на отрезок								Z-преобразование							
	Линейная регрессия		Линейная регрессия с L1 регуляризацией		Линейная регрессия с L2 регуляризацией		Случайный лес		Линейная регрессия		Линейная регрессия с L1 регуляризацией		Линейная регрессия с L2 регуляризацией		Случайный лес	
	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %	Признак	Значимость, %
S	Ca	12,26	Si	16,56	Cl	11,54	Ca	20,47	Si	13,75	Si	16,33	K	16,03	Ca	27,58
	Si	12,24	Ca	13,94	Sr	10,89	Mo когерентный	9,84	Ca	11,37	Ca	11,70	Ca	13,58	Sr	11,05
	K	11,44	K	11,80	Ca	10,07	K	9,25	Sr	9,70	Sr	10,11	Si	11,75	K	9,56
	Sr	10,55	Sr	11,37	K	8,74	площадь фона	9,06	K	8,57	K	9,62	Sr	10,43	Mo	9,08
	площадь фона	8,10	площадь фона	9,62	площадь фона	7,67	Mo	8,77	S	7,55	S	7,62	площадь фона	8,60	Cl	8,41
Сушка	K	17,08	K	18,30	K	18,34	P	11,49	Ca	15,57	Ca	15,68	K	18,23	Ti	11,73
	Ca	14,70	Ca	14,64	Ca	14,80	средняя яркость	8,70	K	13,88	K	15,38	Ca	13,24	Ca	8,76
	площадь фона	11,62	площадь фона	11,72	Ti	10,02	Cl	8,26	площадь фона	9,84	площадь фона	10,29	площадь фона	11,75	Cl	7,34
	Ti	9,76	Ti	9,95	Cl	7,02	количество контуров	8,07	Sr	9,65	Sr	9,25	Mn	9,90	средняя яркость	6,98
	S	8,29	S	8,66	Fe	6,97	K	7,73	Mn	9,24	Ti	9,02	Ti	9,24	P	6,96

Отдельно стоит обратить внимание на различие показателей качества работы регрессии с различной регуляризацией. Как правило, показатели качества L2 регуляризации выше за счет того, что они являются более устойчивыми (обладают меньшим разбросом качества результатов). Интересно так же отметить, что для классификации по химическому составу интенсивности самих определяемых элементов зачастую оказываются не так значимы, как интенсивности линий более тяжелых атомов. Данное явление вызвано сложным составом матрицы исследуемых объектов и широким диапазоном определяемых концентраций и типов исследуемых удобрений. При учете матричных влияний классификатор настраивается на наиболее влияющие элементы: обладающие большим коэффициентом поглощения рентгеновского излучения или являющиеся показателями технологического процесса (как например кремний или кальций). Тем самым нивелируя матричное влияние и обеспечивая универсальность классификации для широкого круга объектов, в том числе произведенных на разных предприятиях по разным технологическим схемам. Возможно именно этим эффектом вызваны неудачи внедрения метода ЭД РФА на производстве сложных фосфорсодержащих удобрений.

Далее, с использованием обобщенных данных, проведена регрессия по множеству компонент для определения химического состава исследованных объектов по основным питательным элементам. Регрессия, в отличие от классификации, позволяет провести непрерывную оценку данных в широком диапазоне значений. Результат работы регрессии по основным питательным элементам (при нормализации данных на отрезок) приведен в таблице 6.12. В качестве метрик качества выступают абсолютное и относительное отклонение тестовых данных от предсказанных значений и коэффициент корреляции полученной прямой. Все метрики рассчитывались по стратегии кросс-валидации, приведенной в главе 3 настоящей работы. Так же стоит отметить, что для регрессии не обнаружено значимого влияния типа предварительной нормализации данных.

Таблица 6.12. Регрессия по основным питательным элементам и сере

Алгоритм	Показатель и диапазон значений, масс. %	абсолютное отклонение		СКО		R^2	
		Среднее	СКО	Среднее	СКО	Среднее	СКО
Линейная регрессия	N [0; 16]	- 0,3901	0,03	0,3750	0,1	0,9878	0,004
Линейная регрессия с L1 регуляризацией		-0,5980	0,06	-0,7807	0,2	0,9704	0,007
Линейная регрессия с L2 регуляризацией		-0,3644	0,03	-0,3266	0,09	0,9878	0,003
Логистическая регрессия	P [15; 52]	-1,1247	0,09	-3,3434	1	0,9800	0,007
Логистическая регрессия с L1 регуляризацией		-1,3247	0,1	-4,8557	1	0,9710	0,009
Логистическая регрессия с L2 регуляризацией		-1,1177	0,06	-3,4988	1	0,9770	0,009
Логистическая регрессия	K [0; 20]	-0,2906	0,03	-0,2368	0,09	0,9961	0,001
Логистическая регрессия с L1 регуляризацией		-0,4255	0,06	-0,4441	0,1	0,9926	0,002
Логистическая регрессия с L2 регуляризацией		-0,2876	0,03	-0,2361	0,09	0,9961	0,001
Логистическая регрессия	S [0; 20]	-0,7002	0,06	-1,1544	0,3	0,9817	0,005
Логистическая регрессия с L1 регуляризацией		-1,0376	0,08	-2,6552	0,5	0,9579	0,009
Логистическая регрессия с L2 регуляризацией		-0,6983	0,06	-1,1519	0,3	0,9817	0,005

В случае регрессии наблюдается ухудшение предсказательной способности в сравнении с классификацией. Данный эффект в первую очередь обусловлен относительно малым разнообразием базы данных (менее 10 объектов с принципиально различным химическим составом). Тем не менее достигнутый уровень абсолютного отклонения не превышающий 1,5 масс. % является неплохим результатом для указанных диапазонов концентраций. Как и предполагалось ранее,

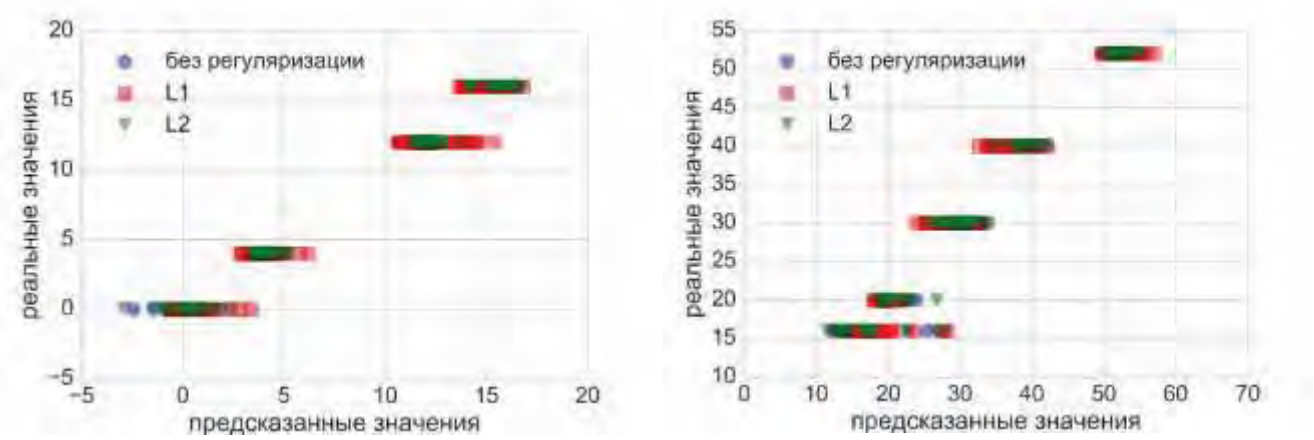
в данных присутствует большое число корреляций между признаками и регрессия с регуляризацией улучшает показатели метрик качества. Интересно отметить, что нормировка на среднее и дисперсию значимо не изменяет данные метрики качества. Значимость признаков приведена в таблице 6.13.

Таблица 6.13. Относительный вклад каждого коэффициента перед признаком в регрессию

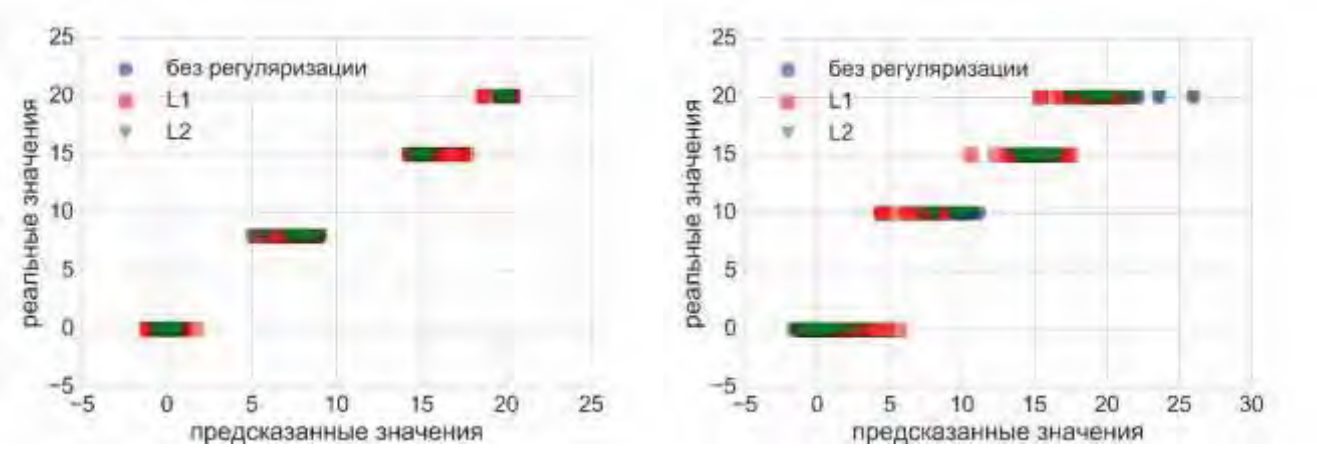
Фактор	**регрессия**		**Линейная регрессия с L1 регуляризацией (Lasso)***		**Линейная регрессия с L2 регуляризацией (Ridge)**	
	Признак	**Значимость, %**	**Признак**	**Значимость, %**	**Признак**	**Значимость, %**
N	K	20,19	Sr	17,36	Ca	23,37
	Sr	15,17	площадь фона	14,33	Sr	19,52
	Ca	14,77	S	9,41	Si	11,35
	S	10,26	Mo когерентный	9,19	K	10,82
	P	7,13	Ti	8,61	S	9,99
P	Sr	20,16	площадь фона	25,94	Sr	22,63
	K	14,27	Sr	16,41	Cl	17,19
	P	11,36	Cl	13,24	K	11,06
	площадь фона	9,37	S	12,74	Mn	9,79
	S	8,81	Mn	5,89	площадь фона	6,63
K	площадь фона	25,44	площадь фона	40,12	площадь фона	27,22
	Ca	12,44	Mo	12,25	Ca	12,25
	Mo	9,76	Mo когерентный	8,38	Mo	10,90
	P	9,30	Sr	7,70	K	9,41
	K	8,86	Ti	6,74	P	8,63
S	K	17,46	Sr	17,18	K	17,49
	Ca	16,17	площадь фона	13,53	Ca	17,01
	площадь фона	15,72	Ti	12,88	площадь фона	15,68
	Ti	10,94	Mo когерентный	11,96	Ti	11,37
	Mo	8,48	Cl	9,56	Mo	8,97

Действительно, значимость признаков в регрессии без регуляризации более равномерно распределена по выделенным свойствам. С другой стороны, для L2 регуляризации (как правило показывает наилучший результат) наблюдается больший размах в значимости признаков. Так же стоит отметить, что значимость оптических свойств для регрессионной прямой не превышает 8 %. Это может быть обусловлено предварительной нормировкой данных, когда дисперсия аналитического сигнала от фракции не так значима для результата (поскольку является не смещенной). Данное предположение подтверждается наихудшим результатом для определения концентрации фосфора, поскольку дисперсия от фракции наиболее сильно сказывается на его аналитическом сигнале (самый легкий элемент из явно определяемых). С другой стороны, для серы дисперсия отчасти нивелируется сигналом фосфора, что положительно сказывается на метриках качества.

Далее приведены графики зависимости предсказанных значений концентраций элементов от их реального содержания в марке (рисунок 6.6). Стоит добавить, что свой вклад в дисперсию вносит как вариация запрессованных фракций, так и варьирование реального химического состава в рамках одной марки.



а) б)



в) г)

Рисунок 6.6. Зависимость предсказанных значений от марки объекта. а) – азот, б) – фосфор, в) – калий, г) – сера.

Интересно отметить, что для регрессии без регуляризации наблюдается отчасти нелинейная зависимость, по виду похожая на кубическую. С другой стороны, L2 регуляризация эффективно устраняет данное явление.

Таким образом, с использованием обобщенных данных становиться возможно автоматически классифицировать исследуемые объекты по фракции, марки и типу удобрения с точностью не менее 98 % и предсказывать концентрации элементов в широких диапазонах с абсолютным отклонением не более 1,5 масс. %, что является отличным результатами для технологического контроля. При этом время получения информации о физических и химических свойствах неизвестного объекта занимает менее 10 минут. Остается сделать анализ более наглядным.

6.4 Кластеризация и визуальное представление данных

Как уже упоминалась ранее, классификация заложена в основу построения классификации без «обучающей» выборки – когда не известны значения ответа для исследуемых данных. Методы этой группы особенно актуальны при общем представлении всей совокупности данных, оценки качества перехода промышленности с одной марки на другую, быстрой оценки мер близости различных объектов по совокупности свойств.

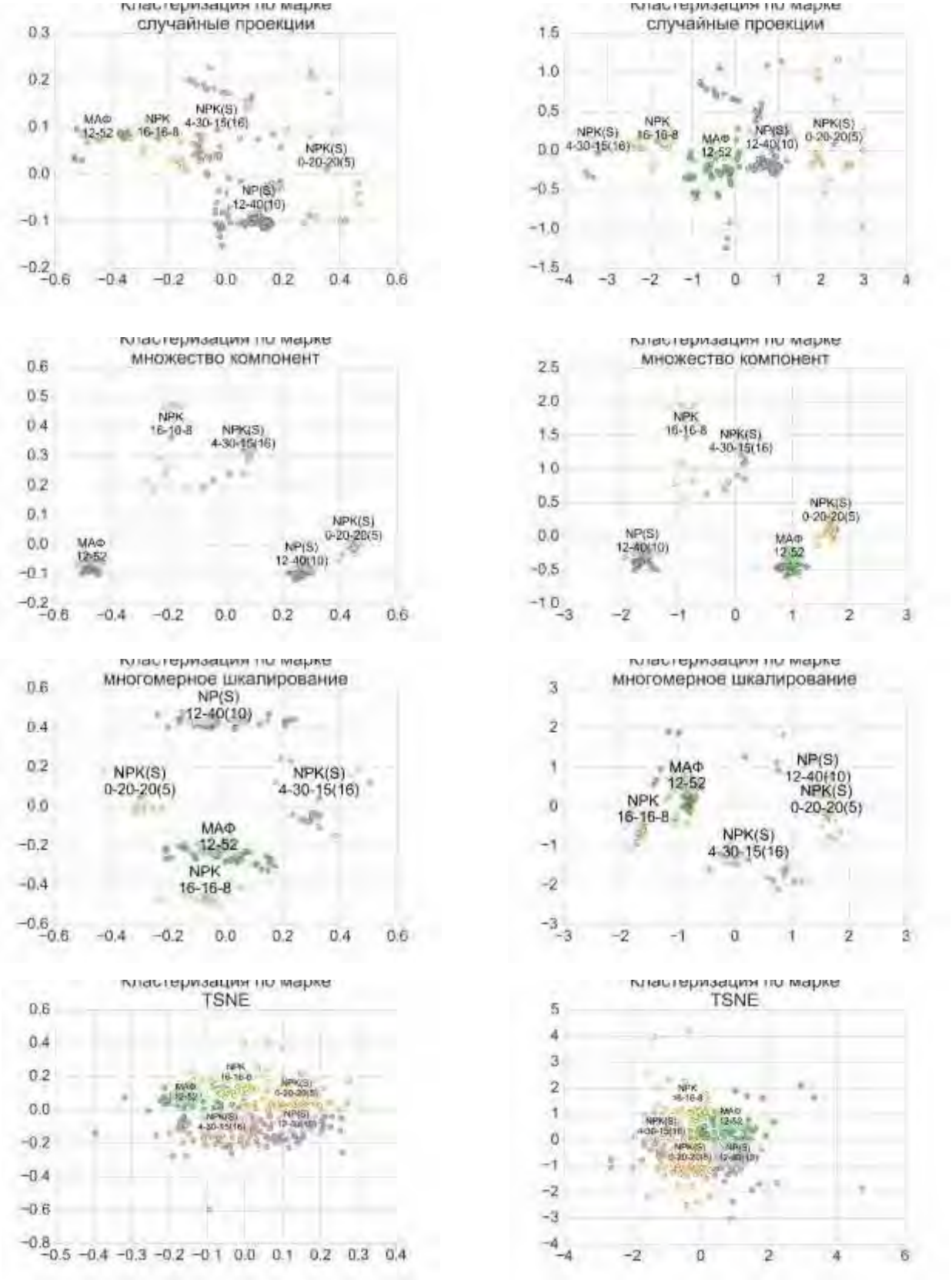
В составленном нами обобщенном наборе данных присутствует порядка 600 объектов с 35 признаками для каждого. Наряду с проведенным ранее отбором признаков по классификации и регрессии с регуляризацией, существуют специализированные методы, позволяющие сократить размерность на основании более сложных механизмов объединения параметров – так называемое «понижение размерности» (глава 3 настоящей работы). При этом понижение размерности позволяет так или иначе учесть все свойства объектов, присутствующие в базе данных. Так, для совокупности исследуемых объектов рассмотрены следующие алгоритмы понижения размерности данных:

- проекция случайных компонент (random PCA, линейное преобразование);
- проекция на главные компоненты (PCA, линейное преобразование);
- многомерное шкалирование (MDS, нелинейное преобразование);
- TSNE (t-распределенных стохастических соседних вложений, нелинейное преобразование).

Наряду с линейными методами понижения размерности данных, описанных в главе 3 настоящей работы приведены некоторые нелинейные алгоритмы, доступные в языке программирования Python 2.7. Обсуждение математического аппарата нелинейных алгоритмов понижения размерности выходит за рамки данной работы.

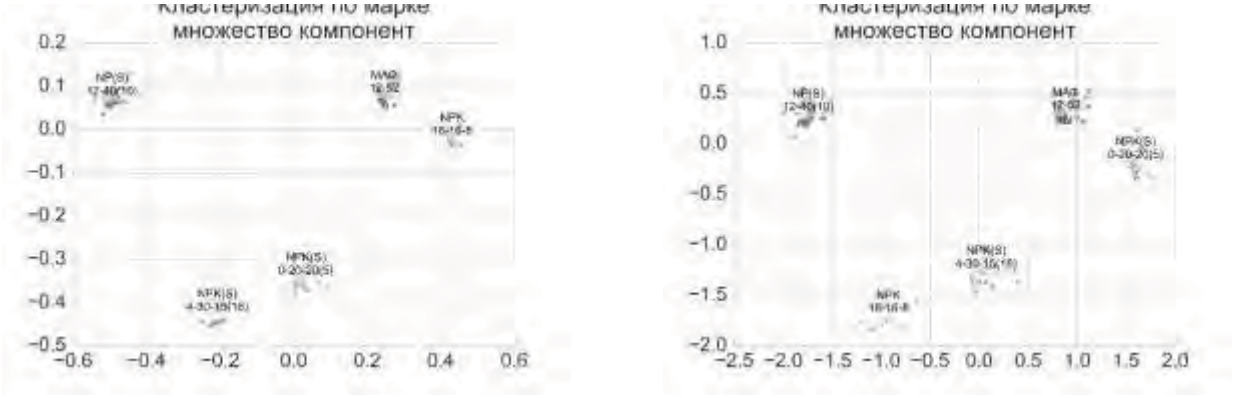
С использованием описанных алгоритмов проведено понижение размерности до двум компонентам с различной нормировкой полученных данных: на отрезок и Z-преобразованием (рисунок 6.7).

Дополнительно, с использованием наиболее интерпретируемого и достаточно точного алгоритма проекции на главные компоненты, проведена кластеризация по порошкам (рисунок 6.8) и по гранулам объектов (рисунок 6.9).



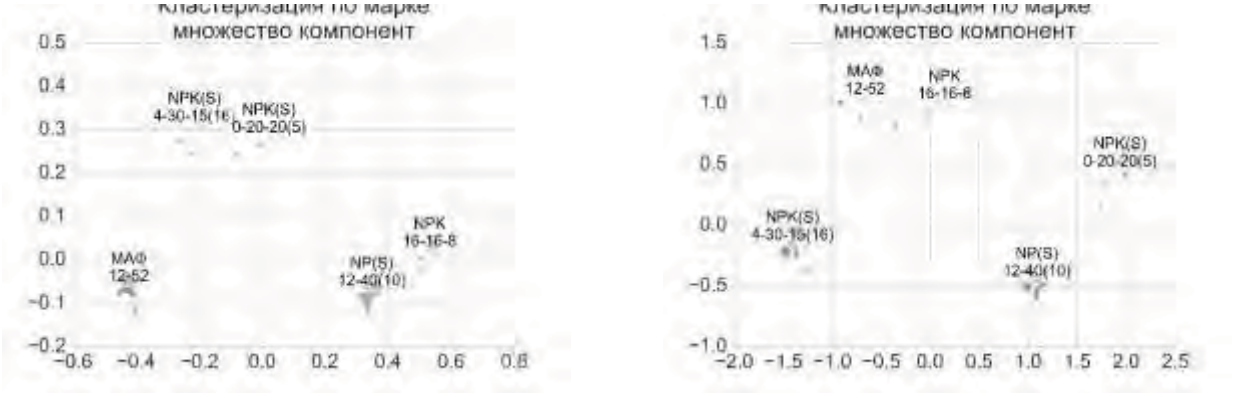
а) б)

Рисунок 6.7. Кластеризация данных. а) – нормирование на отрезок, б) – Z-преобразование



а) б)

Рисунок 6.8. Кластеризация порошков. а) – нормирование на отрезок, б) – Z-преобразование



а) б)

Рисунок 6.9. Кластеризация гранул. а) – нормирование на отрезок, б) – Z-преобразование

По представленным данным видно отличное разделение типа и марки удобрений в пространстве двух признаков с использованием алгоритма проекции на главные компоненты как для общего случая, так и для разных фракций в отдельности. Так же неплохие результаты демонстрирует нелинейный алгоритм многомерного шкалирования, однако интерпретировать нелинейные результаты гораздо сложнее. Следует отметить, что, как и для кластеризация, нормирование на отрезок дает лучшее результаты. То же подтверждается точностью кластеризации данных, нормированных на отрезок, оцененной по механизму k-средних согласно стратегии кросс-валидации, описанной в главе 3 настоящей работы (таблица 6.14).

Таблица 6.14. Значение метрик качества кластеризации мерок удобрений с параметром устойчивости F-меры.

Тип данных	Нормировка*	Понижение размерности до 2 компонент	Точность	Полнота	F-мера	СКО F1, %
Все фракции	1	Случайные проекции	0,8502	0,8211	0,8152	0,16
	2		0,8298	0,8565	0,8270	0,21
	1	PCA	0,9714	0,9571	0,9619	0,076
	2		0,9528	0,9233	0,9348	0,13
	1	MDS	0,9583	0,9281	0,9408	0,12
	2		0,9732	0,9577	0,9634	0,073
	1	TSNE	0,6637	0,6167	0,6185	0,28
	2		0,7863	0,7747	0,7575	0,25
Только порошки	1	PCA	0,9314	0,8869	0,9045	0,15
	2		0,9262	0,8817	0,8986	0,16
	1	MDS	0,9267	0,8805	0,8984	0,16
	2		0,9000	0,9250	0,9000	0,13
Только гранулы	1	PCA	0,9613	0,9474	0,9523	0,06
	2		0,9425	0,9142	0,9215	0,12
	1	MDS	0,9463	0,9182	0,9338	0,10
	2		0,9621	0,9475	0,9524	0,05

* 1 – нормировка на отрезок, 2 – нормировка на дисперсию (Z-преобразование)

По приведенным данным наблюдается интересная закономерность – нормировка на отрезок в целом лучше работает для линейных алгоритмов, в то время как z-преобразование для нелинейных алгоритмов.

Таким образом доказана возможность точного и наглядного представления исследуемых объектов на плоскости по совокупности свойств.

По полученной карте с использованием физических и химических параметров объекта становиться возможным не только визуализировать объекты, но и рассчитать расстояние между ним. Оптимальные расстояния между классами, можно рассчитать как прямые линии между центрами классов с использованием Евклидовой метрики (согласно главе 3 настоящей работы). Значение центров классов объектов приведены в таблицах 6.15 и 6.16. Для сравнения используются различные варианты расчета центра класса.

Таблица 6.15. Центры классов объектов по MDS.

Способ центрирования	Тип	Марка	Значение первой компоненты	Значение второй компоненты
Арифметическое среднее	МАФ	12-52	1,399	1,1550
	NP(S)	12-40(10)	-1,175	-1,029
	NPK(S)	4-30-15(16)	0,2171	-0,1861
	NPK	16-16-8	0,5883	0,6503
	NPK(S)	0-20-20(5)	-0,4862	-0,6315
К-средних	МАФ	12-52	0,5197	0,7009
	NP(S)	12-40(10)	-1,175	-1,029
	NPK(S)	4-30-15(16)	1,397	1,098
	NPK	16-16-8	-1,155	0,6691
	NPK(S)	0-20-20(5)	0,9373	-1,0877

Таблица 6.16. Центры классов объектов по PCA.

Способ центрирования	Тип	Марка	Значение первой компоненты	Значение второй компоненты
Арифметическое среднее	МАФ	12-52	0,2696	-0,07799
	NP(S)	12-40(10)	-0,4700	-0,06162
	NPK(S)	4-30-15(16)	0,4424	0,004761
	NPK	16-16-8	-0,0704	0,3143
	NPK(S)	0-20-20(5)	0,0264	0,3105
К-средних	МАФ	12-52	0,06648	0,3224
	NP(S)	12-40(10)	0,2702	-0,07785
	NPK(S)	4-30-15(16)	-0,4700	-0,06162
	NPK	16-16-8	0,4516	0,0101
	NPK(S)	0-20-20(5)	-0,1832	0,3683

Таким образом, по информации, полученной от программно-аппаратного комплекса становиться возможно описать конкретный исследуемый объект по совокупности физико-химических свойств и представить их совокупность на плоскости. Предложенные алгоритмы позволяют оценить оптимальность процесса перехода промышленного комплекса с производства одной марки на другую. Чем ближе полученные значения располагаются к прямой, соединяющей центры классов, тем лучше происходит переход.

7 Программная реализация алгоритмов и проведение опытно-промышленных испытаний

В данной главе описывает алгоритм работы программного обеспечения (ПО) созданного аналитического комплекса, а также проведение его различных опытно-промышленных испытаний в рамках предприятий холдинга «ФосАгро».

7.1 Программное обеспечение

Для реализации описанных алгоритмов классификации, регрессии и кластеризации создана и занесена в государственный реестр программа для ЭВМ «DSpectra». Программное обеспечение (ПО) написано на языке программирования Python 2.7, является кроссплатформенным (ОС семейства Linux и Windows) и содержит в себе порядка 10000 строк кода. Программное обеспечение разработано для обработки и преобразования оптической и спектральной информации, полученной от оптического регистратора собственной конструкции и энергодисперсионных рентгеновских спектрометров любого производителя при наличии возможности экспорта спектра в виде вектора интенсивностей или матрицы энергия-интенсивность в любой файловый формат. ПО позволяет проводить:

1. Математическую обработку оптической и спектральной информации согласно главе 3 настоящей работы:
    a. Сглаживание сигнала;
    b. Приведения исходного спектра к дифференциальной, нормализованной и частотной форме.
2. Получение аналитической информации:
    a. Расчет аномалий на карте поверхности изображения;
    b. Ручное выделение и маркировка характеристических линий;
    c. Расчет интенсивности, площади и отношения сигнал-шум для пика;

d. Вычисление и нивелирование базовой линии;

e. Автоматическое проведение всех вышеперечисленных стадий.

3. Методическую работу с аналитической информацией:

a. Создание библиотеки полученной информации;

b. Построение матриц «объекты-признаки»;

c. Классификация и множественная регрессия согласно главе 3 настоящей работы;

d. Создание пользовательских методов регрессионного анализа;

e. Сохранение пользовательских схем анализа.

4. Возможность поточной обработки информации по каждой представленной стадии.

5. Экспорт полученной информации в виде *.txt, *.csv, *.xls файлов.

6. Построение отчетов о проделанной работе.

Структура ПО, основанная на ранее выбранных и оптимизированных алгоритмах работы с физико-химическими признаками удобрений, приведена на рисунке 7.1. Общий вид ПО приведен на рисунке 7.2.

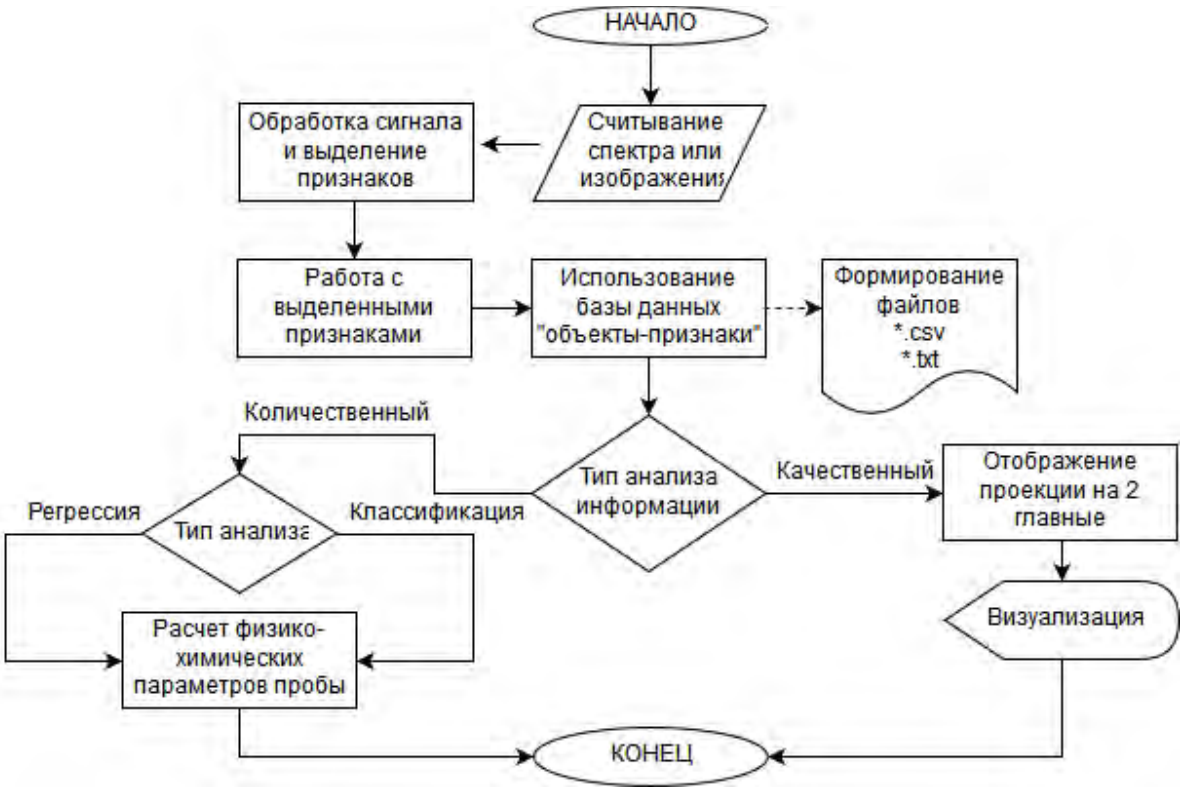


Рисунок 7.1. Структура ПО «DSpectra».

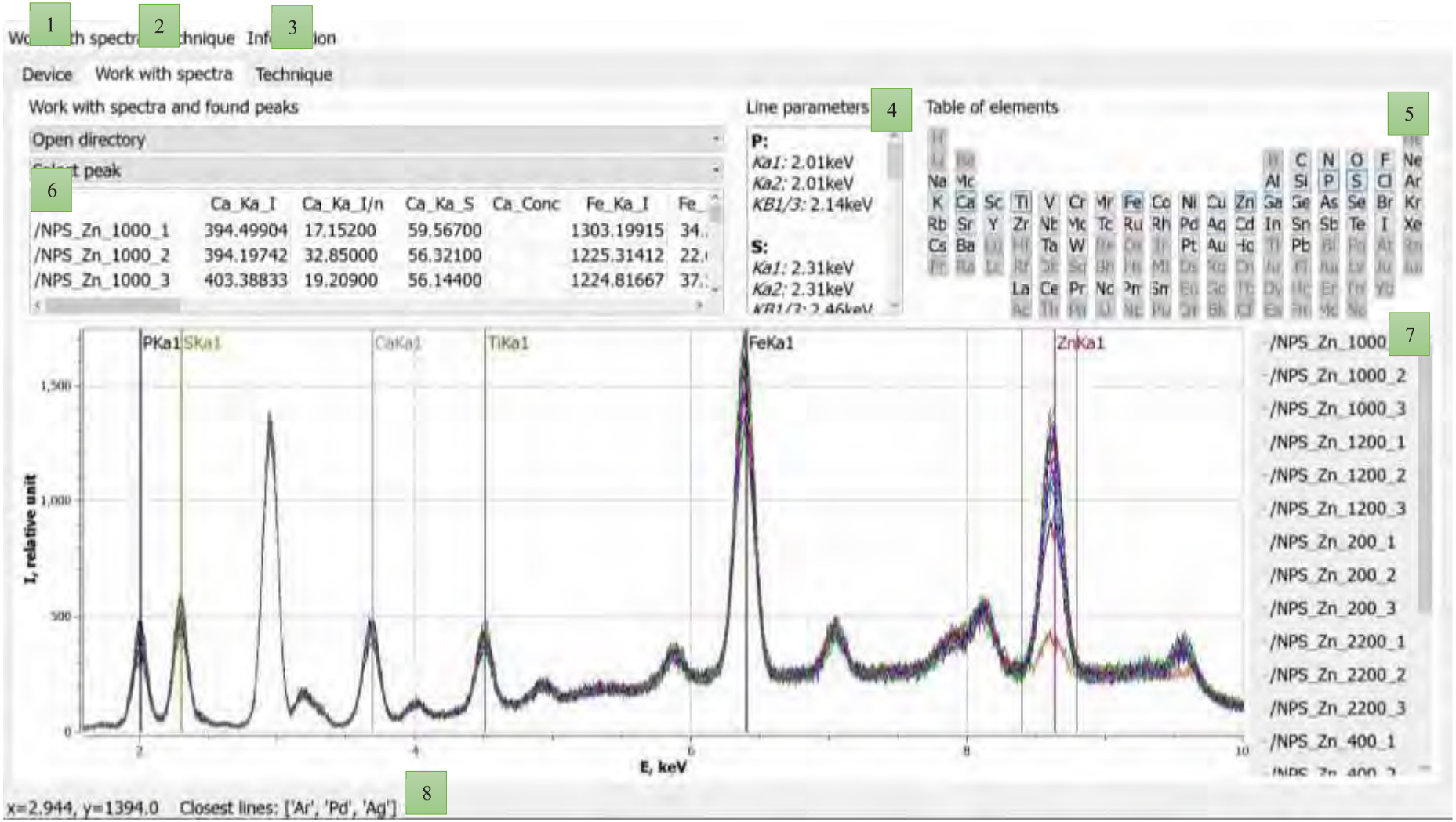


Рисунок 7.2. Общий вид разработанного ПО. 1, 2 и 3 – вкладки работы с оборудованием, спектрами и признаками соответственно, 4 – окно информации по характеристическим линиям, 5 – окно выбора элементов, 6 – данные по физико-химическим свойствам спектров, 7 – список открытых спектров, 8 – информация с курсора.

Дополнительной особенностью программного обеспечения является возможность отслеживания качества работы лаборанта при подготовке проб к анализу по значению максимальной фракции и качеству классификации типа удобрения. Программная реализация основных алгоритмов, использованных в работе и реализованных в ПО приведена в приложении А.

7.2 Решение аналитических задач производства минеральных удобрений

С использованием описанной рентгенооптической установки реализован набор актуальных аналитических задач производства сложных фосфорсодержащих удобрений.

7.2.1 Оптимизация пробоподготовки гранулированных продуктов

Ключевой особенностью аналитического контроля на промышленном производстве является постоянное стремление к автоматизации и сокращению времени анализа. Для достижения данных целей требуется оптимизация пробоподготовки промышленных объектов, которая является самой затратной по времени и по величине погрешности стадией практически любого химического анализа.

Рассматриваемые в данной работе объекты объединяет одно общее свойство – это технические пробы. Данные пробы отличаются достаточной изменчивостью состава (порядка 1 - 5 %) по основным элементам даже в рамках одной марки. В нашем случае нужно подобрать такую процедуру подготовки пробы, чтобы она:

- занимала менее 15 минут;
- была воспроизводима;
- давала устойчивую пробу, пригодную для хранения и повторного анализа;
- была универсальной и однотипной для широкого класса объектов (например апатит, удобрения и т.д.);

- была просто реализуемой за рамками одного завода – например, когда требуется сопровождать выпуск экспериментальной партии удобрений или развернуть систему РФ-анализа на другой производственной площадке.

Как следует из вышеприведенных требований (расположенных в порядке убывания значимости), у нас практически не остается времени на сложные операции с пробой, такие как концентрирование или перевод пробы в другую форму (например растворение). То же относиться и к истиранию пробы до минимально требуемой для РФА крупности в 70 мкм и ситовому контролю крупности частиц пробы.

В качестве основного объекта для исследования выбрано NPKS удобрение марки 4-30-16(15), выпускаемое на предприятии АО «ФосАгро-Череповец». Данное удобрение обладает наиболее разнообразным элементным составом и, как следствие, наиболее неоднородное из всех рассматриваемых объектов, что подтверждается данными сканирующего ЭД-РФА (рисунок 5.22).

Исследуемый объект готовился для анализа несколькими способами согласно главе 4 настоящей работы (таблица 4.2), а затем прессовался в виде «сэндвич-структуры» на подложке из борной кислоты с усилием порядка 260 бар.

Каждый тип объекта готовили в 10 параллельных излучателях. Каждый излучатель измеряли на РФ-спектрометре согласно условиям, подобранным в пункте 5.3.1 данной работы.

Для каждого исследуемого объекта оценивали набор свойств:

1. фракция (100, 500 мкм и гранулы);
2. предварительная сушка (0 – нет, 1 – да);
3. максимальная интенсивность общей фоновой линии;
4. максимальная интенсивность характеристических линий элементов (таблица 7.1) и их энергия.

Таблица 7.1. Индикаторные характеристические линии элементов.

Элемент	Энергия, кЭв
P	2,0
S	2,3
K	3,3
Ca	3,7
Mo	17,4
Mo когерентное рассеяние	16,5

Результат расчета параметров проб для разных фракций объекта с помощью оптимизированного алгоритма согласно главе 5 настоящей работы приведен на рисунке 7.3.

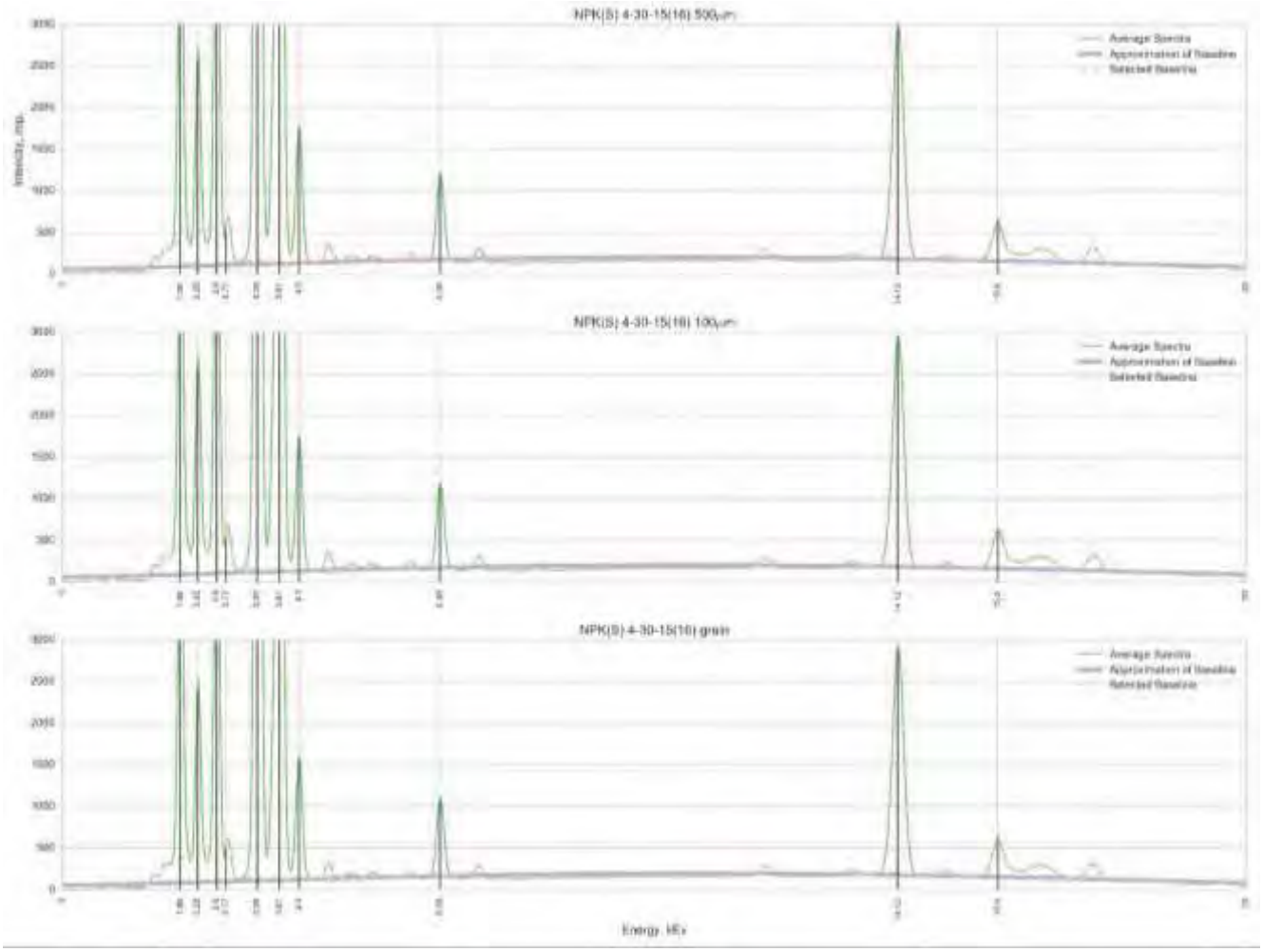


Рисунок 7.3. Результат автоматического расчета базовой линии и интенсивностей характеристических линий для усредненных спектров исследуемых типов объектов (прессованных гранул, порошка < 500 мкм и порошка <100 мкм).

В результате получена матрица «объекты-признаки» размером 108 х 35. В качества признаков выступают максимальная интенсивность аппроксимированной

фоновой линии (фон) и энергии найденных характеристических линий (1,75 – Si Kα, 2,0 – P Kα, 3,65 – Ca Kα и т.д.). На рисунке 7.4 приведен обзор взаимосвязей «признак-признак» для нормированных на отрезок интенсивностей характеристических линий и фоновой линии:

$$I_{norm} = \frac{I - I_{mean}}{I_{max} - I_{min}}$$

где: I_{norm} – нормированная, I_{mean} – средняя, I_{max} – максимальная и I_{min} – минимальная интенсивности признаков.

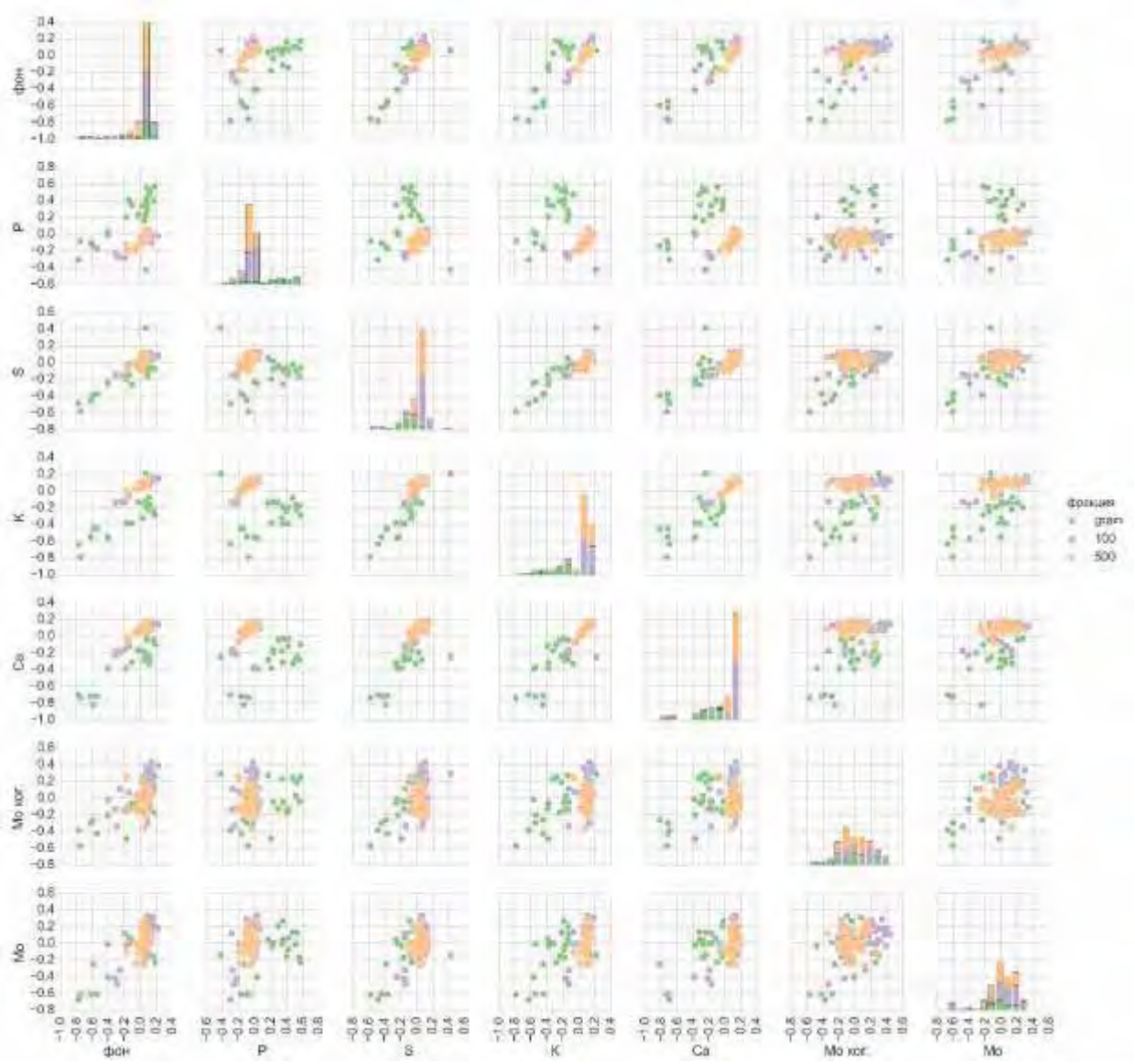


Рисунок 7.4. Взаимосвязь признаков спектра для классификации крупности частиц

Приведенное представление используется для выявления статистических особенностей исследуемых объектов по целевой переменной (в данном случае – фракция). Можно обратить внимание на различие дисперсий и разницу в средних значениях аналитических сигналов в зависимости от фракции. Наиболее широким распределением предсказуемо обладают прессованные гранулы, в то время как распределения порошков достаточно похожи по форме и ширине, что может говорить о потенциально равнозначном использовании порошков фракций 100 и 500 мкм. Однако наблюдается определённое влияние фракции на максимальную интенсивность фона и интенсивности линий средних (по атомной массе) элементов (калий).

На следующем рисунке (рисунок 7.5) приведена взаимосвязь признаков спектра для классификации по влажности объектов.

По полученным данным не выявлено сильной информативности отдельных признаков для классификации образцов по влажности. Это может говорить о том, что присутствующая в объектах влага значимо не влияет на результаты анализа.

На рисунке 7.6 приведена карта линейных корреляций между признаками. По полученной карте можно отметить, что сильная линейная корреляция наблюдается для признаков «фракция – средние элементы», в особенности калий. Для признаков «сушка – энергии элементов» значимой корреляции не прослеживается.

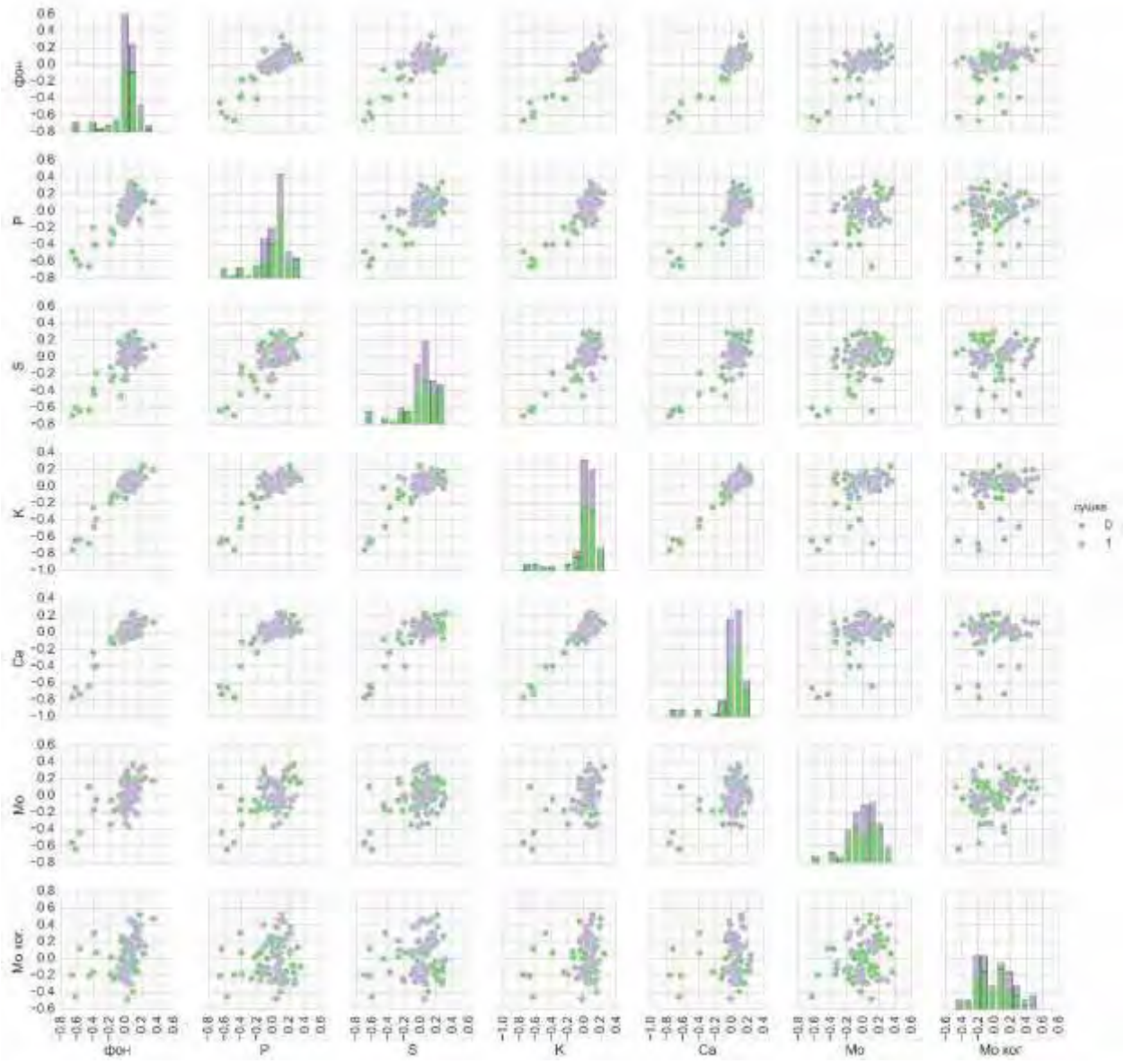


Рисунок 7.5. Взаимосвязь признаков спектра для классификации по влажности.

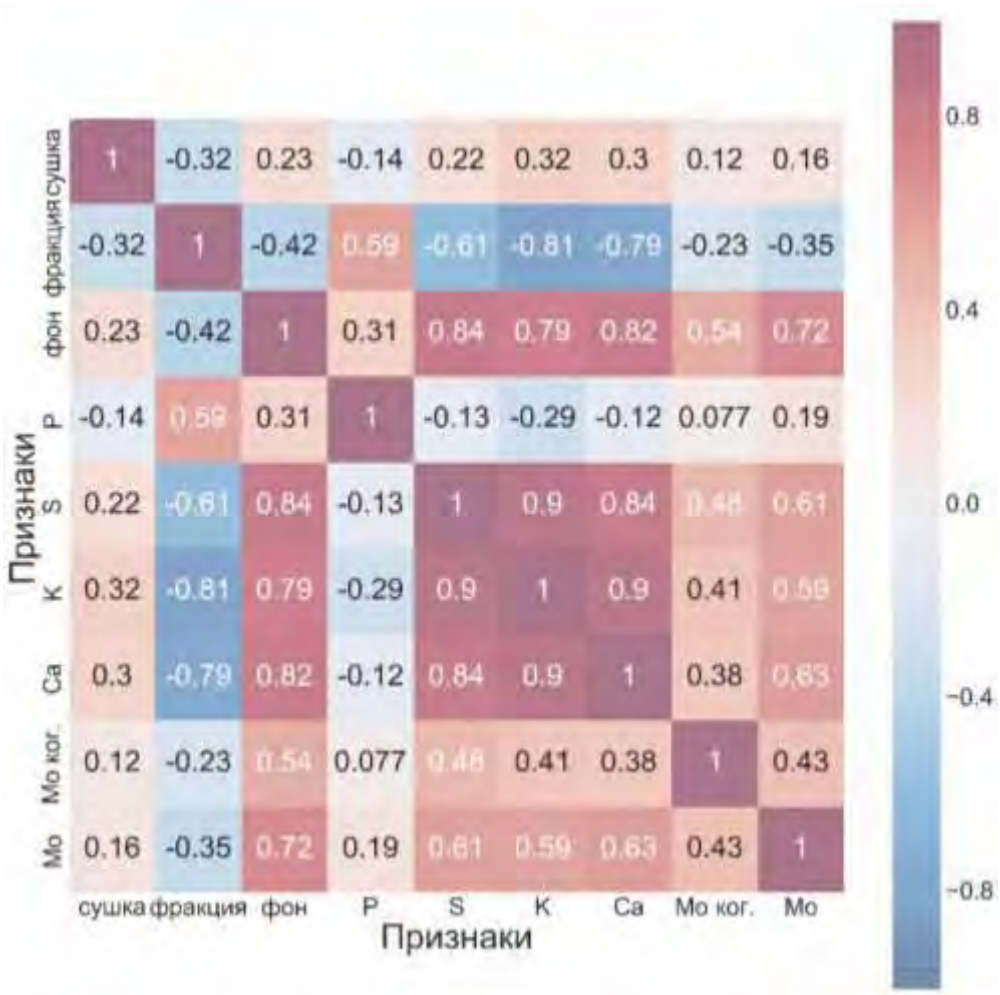


Рисунок 7.6. Карта линейных корреляций между признаками спектра.

Таким образом, наиболее подверженными влиянию фракции элементами являются характеристические линий средних (по атомной массе) элементов, в особенности калия. Рассчитанные статистические показатели для каждого типа пробоподготовки исследуемого объекта позволяют рекомендовать для проведения количественного РФ-анализа удобрений истирание до фракции 500 мкм с предварительным высушиванием. С другой стороны, с учетом ограничения по времени проведения технологического контроля, от предварительной сушки пришлось отказаться, заменив ее большим количеством параллельных измерений.

7.2.2 Мониторинг переходного процесса

Используя разработанную в подразделе 6.4 схему понижения размерности данных и визуализации, проведен контроль процесса перехода промышленного

комплекса с производства NP(S) удобрения марки 12-40(10) на NPS(S)+Zn удобрение марки 12-40-5(4)+1. Переход проходил в течении трех суток, за которые было выпущено порядке 1000 тон продукта. Отбор проб проводился каждый час с классическим контролем химического состава по N, P, S и Zn и параллельным РФ анализом порошков фракции 500 мкм. Время РФ анализа (с учетом 2 параллельных проб) составило 10 минут, взамен 240 минут классическими методами.

Поскольку промышленный переход в данном случае заключался во внесении дополнительного количества элементной серы и оксида цинка, новая марка отличается от производимой по двум параметрам и хорошо представима на плоскости (рисунок 7.7).

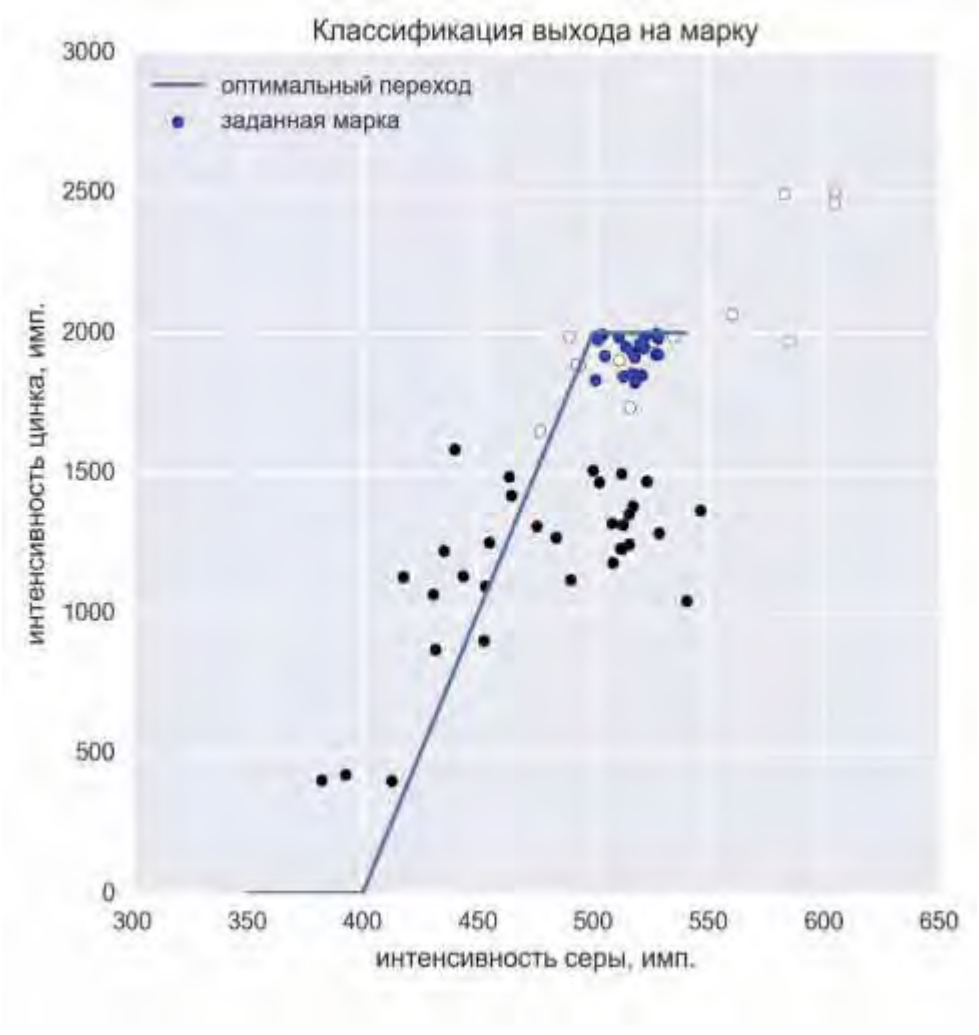
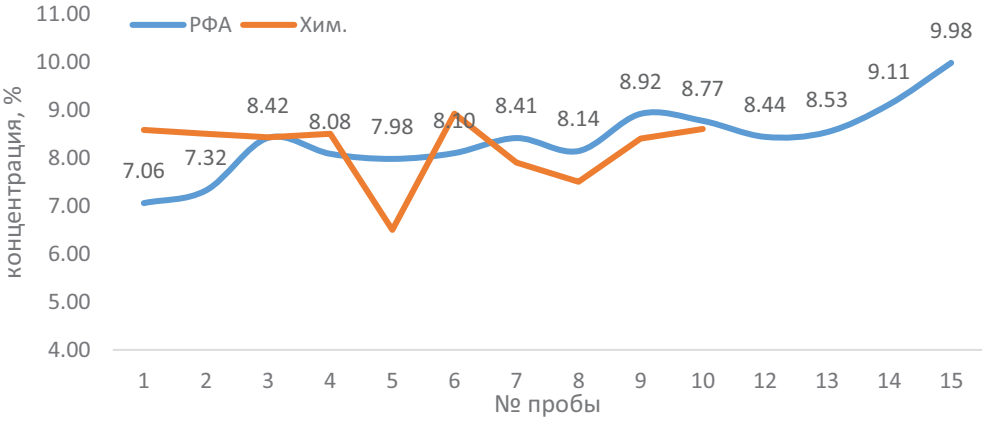


Рисунок 7.7. Визуализация промышленного процесса перехода с марки на марку по двум компонентам (сера и цинк). Черными точками отмечен переходный продукт, белыми – продукт надлежащего или лучшего качества. Синими точками отмечена «идеальная» марка.

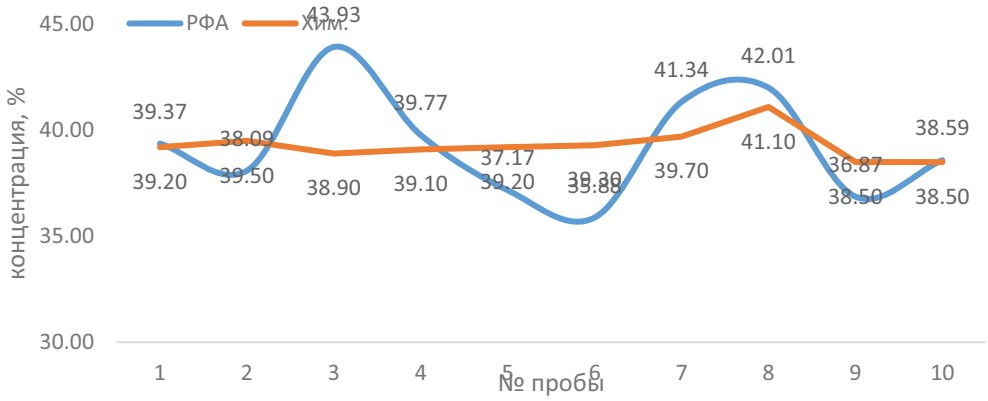
По представленным данным можно обратить внимание, что переход осуществлялся не идеально и с достаточно сильным разбросом относительно прямой оптимального перехода.

Графики колебаний концентраций основных питательных элементов и цинка в сравнении с классическими методами анализа (спектрофотометрия) приведены на рисунке 7.8.

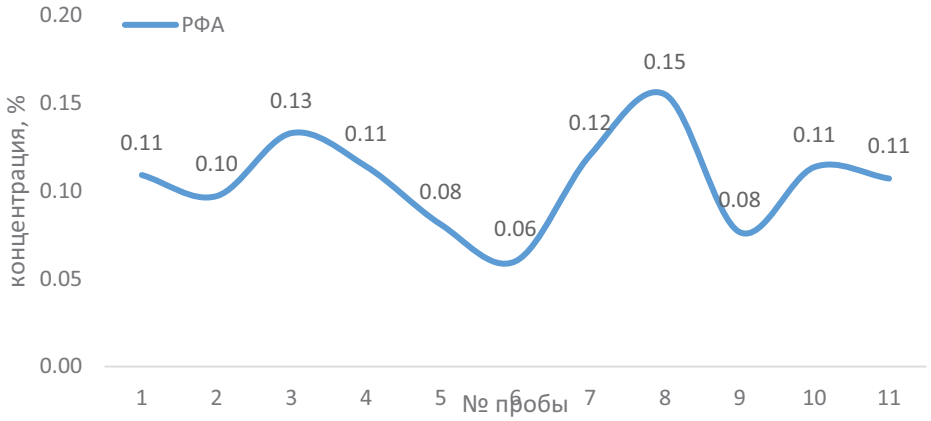
а)



б)



в)



г)

Рисунок 7.8. Сравнение значений РФА метода и химического контроля, используемого на предприятии. Нумерация объектов ведется от начала испытаний: а) – Zn, б) – S, в) – P_2O_5, г) – CaO.

Классификация производимой продукции проводилась с использованием понижения размерности методом главных компонент. По исследованию, проведенному в главе 6 настоящей работы, данный метод обеспечил вторую по качеству кластеризацию и является линейным, что упрощает интерпретацию результатов визуализации. Результат работы алгоритма приведен на рисунке 7.9.

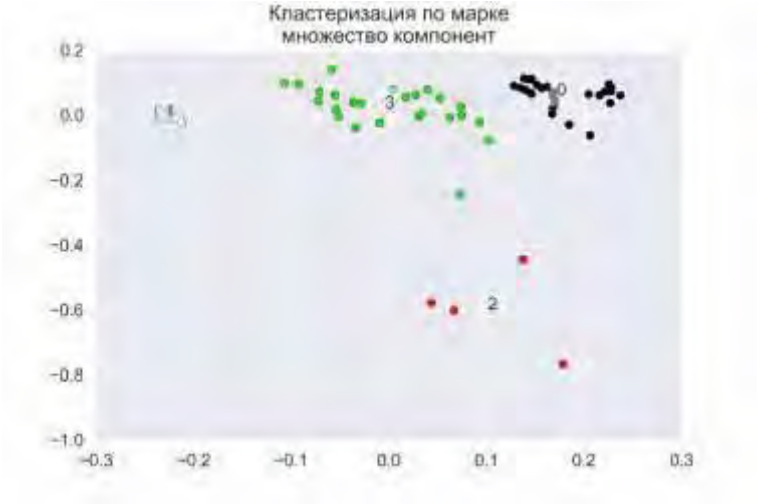


Рисунок 7.9. Пример понижения размерности и автоматической кластеризации переходного процесса. 0 – конечный продукт, 1 – исходный продукт, 2 – значимые выбросы, проблемы в технологии, 3 – процесс перехода.

Алгоритм автоматизирован и сам предсказывает аномалии, начало и конец переходного процесса. Исходя из полученных результатов представляется возможным не только качественная кластеризация и визуализация различных

типов удобрений, но и автоматическое отслеживание качества прохождения промышленных процессов по всей совокупности получаемых физических и химических свойств. При этом время анализа одной пробы не превышают 10 минут.

7.2.3 Решение «нетривиальных» задач контроля качества

Наряду с классическим контролем качества по химическому составу в промышленной практике появляются уникальные показатели, приближающие те или иные физические и химические параметры объектов. Такие параметры так же являются достаточно значимыми, поскольку обеспечивают более понятную интерпретацию промышленного процесса для конечного пользователя. Примером таких показателей являются:

- фактор формы гранул удобрений – величина, оценивающая «сферичность» гранул по которой косвенно судят о качестве производственного процесса;
- количество кондиционера – показатель расхода самого дорогого химического вещества в производстве (как правило высокомолекулярные органические соединения, флуоресцирующие в УФ диапазоне);
- солевой индекс – величина, оценивающая осмотическое давление почвенных растворов после внесения того или иного удобрения, зависит от растворимости удобрений и проводимости их растворов.

Как правило, подобные показатели достаточно просто измерить на практике, однако данный процесс трудоемок и отнимает много времени. Далее приведены примеры расчета описанных полуколичественных показателей с использованием созданной рентгенооптической схемы.

7.2.3.1 Карты качества гранулированных продуктов

Для описания производственного процесса и физических свойств производимой продукции часто оценивают гранулометрический состав и фактор формы удобрений. Предложенная нами схема оптического регистратора позволяет

работать не только с пробами, подготовленными к РФ-анализу, но и с гранулами удобрений напрямую.

Для расчета гранулометрического состава использовались алгоритмы, описанные в пункте 3.2.1 настоящей работы. Исходное изображение гранул с необработанной картой поверхности представлено на рисунке 7.10.

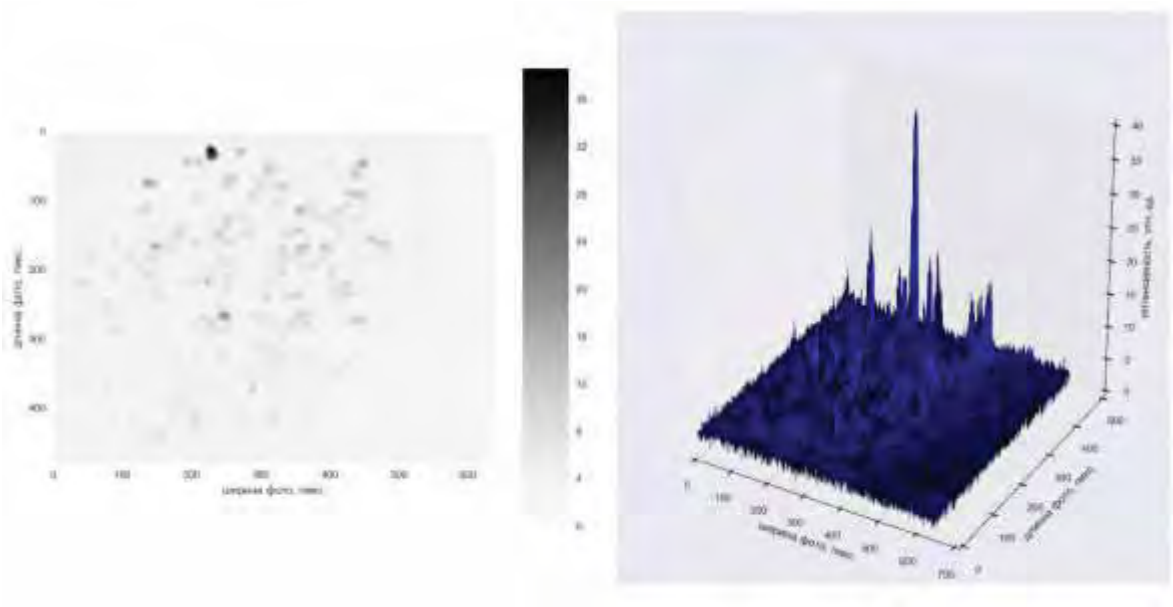


Рисунок 7.10. Исходные изображения гранул с трехмерной картой поверхности.

После стадии дифференцирования, сглаживания и бинаризации карты поверхности проводится выделение гранул и аппроксимация их эллипсами (рисунок 7.11).

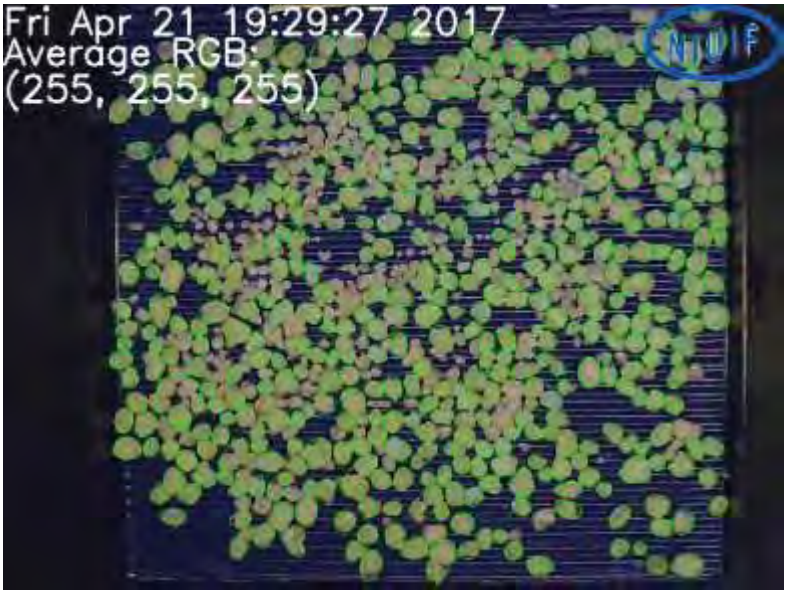


Рисунок 7.11. Обработанная информация

По аппроксимированным эллипсам выделяется длинная ось, короткая ось и цветность гранулы в системе RGB. По данным параметрам рассчитываются такие показатели качества выпускаемой продукции как гранулометрический состав и фактор формы. Полученные данные по определению гранулометрического состава приведены в таблице 7.2 и на рисунке 7.12.

Таблица 7.2. Сравнение показателей гранулометрического состава, полученного различными методами. Продукт МАФ, БФ АО «Апатит».

фракция, мм	**прибор camsizer, %**	**круглые сита, %**	**плетеные сита, %**	**оптический, %**	**абсолютное отклонение оптического метода от, %**		
					camsizer	**круглые сита**	**плетеные сита**
0-1	0,00	0,02	0,03	0,00	0,00	0,02	0,03
1-2	0,27	0,09	0,16	0,00	0,27	0,09	0,16
2-3	20,57	11,05	40,29	27,35	6,78	16,30	12,94
3-4	75,13	81,14	56,12	56,50	18,63	24,64	0,38
4-5	3,93	7,52	3,35	15,31	11,38	7,79	11,96
>5	0,10	0,19	0,05	0,84	0,74	0,65	0,79

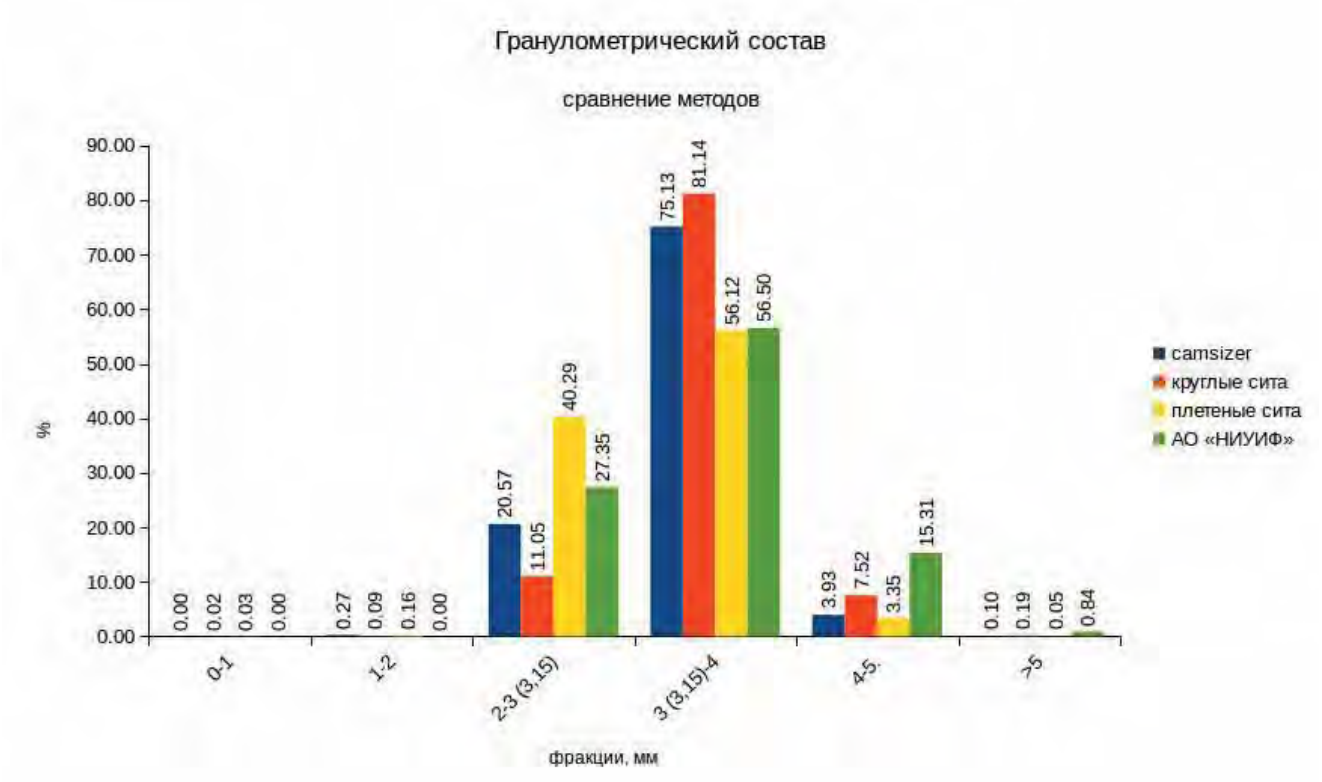
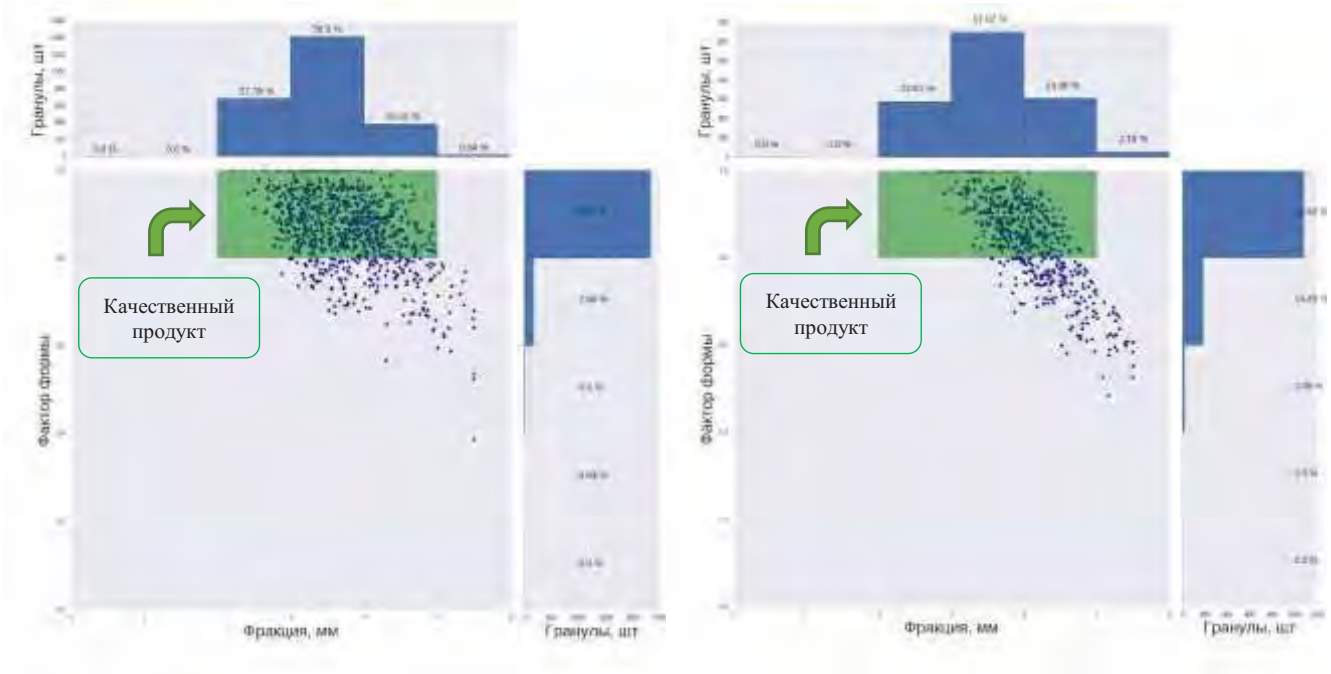


Рисунок 7.12. Гистограммы распределения фракций.

По представленной информации видны значимые расхождения между всеми методами анализа, однако общий вид распределения представлен каждым способом достаточно хорошо и каждый анализ согласуется друг с другом. Предложенный метод оптического контроля хорошо согласуется с данными ситового анализа на плетеных ситах. Это объясняется программной аппроксимацией гранул эллипсами, которые лучше представляют диагональ квадратной ячейки плетеного сита. Завышение в крупных фракциях обусловлены несовершенством программного обеспечения: на сегодняшний день программа определяет некоторые гранулы совместно друг с другом, что завышает долю крупной фракции за счет мелкой.

Используя полученную информацию можно построить «карты качества» выпускаемой продукции. Данные карты представляют из себя график зависимости фактора формы от фракционного состава, позволяющий выделить область качественного продукта по ГОСТ и ТУ для каждого типа и марки удобрения (рисунок 7.13).



а) б)

Рисунок 7.13. Зависимость фактора формы от гран. состава, а) – МАФ, производства БФ АО «Апатит», б) – ДАФК, производства АО «ФосАгро-Череповец».

Видно, что технология производства в случае МАФ дает в целом удовлетворительный результат, в то время как для ДАФК возможно требуется доработка существующей технологии производства.

Разработанные методы оценки позволяют проводить экспрессный и наглядный контроль производственных процессов, а также сделать вывод о пригодности оптического регистратора к работе не только в совокупности с ЭД РФ-анализаторами, но и в качестве самостоятельной стадии контроля производимой продукции.

7.2.3.2 Качество обработки кондиционирующими добавками

Кондиционирующие добавки (к.д.) – самый дорогой реактив, использующийся при производстве минеральных удобрений, расход которого для обработки поверхности гранул подлежит строгому контролю. Степень обработки кондиционирующими добавками определяется также, как и физические свойства гранул удобрений (в предыдущем пункте), однако в качестве источника освещения выступает только УФ-излучатель с длинной волны не более 400 нм. Современные CMOS камеры обладают большей чувствительностью, чем человеческий глаз и УФ-излучения достаточно, чтобы распознать гранулы на фотографии (рисунок 7.14).

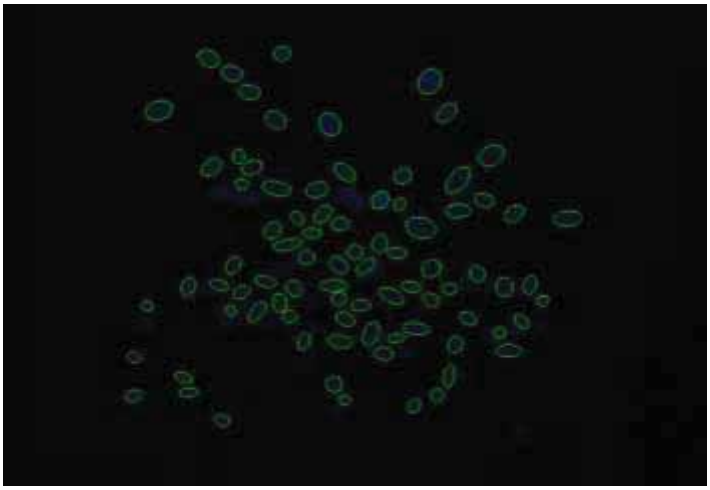


Рисунок 7.14. Изображение распознанных гранул в УФ свете.

Для регрессионного анализа качества обработки к.д. исследован набор продуктов, включающий в себя удобрения МАФ и NPK 15-15-15, производства АО

«ФосАгро-Череповец», обработанные к.д. в лабораторных и промышленных условиях с дозировкой 0,25, 0,3 и 0,4 кг/т. Результаты по каждой стадии математического преобразования приведены на рисунке 7.15.

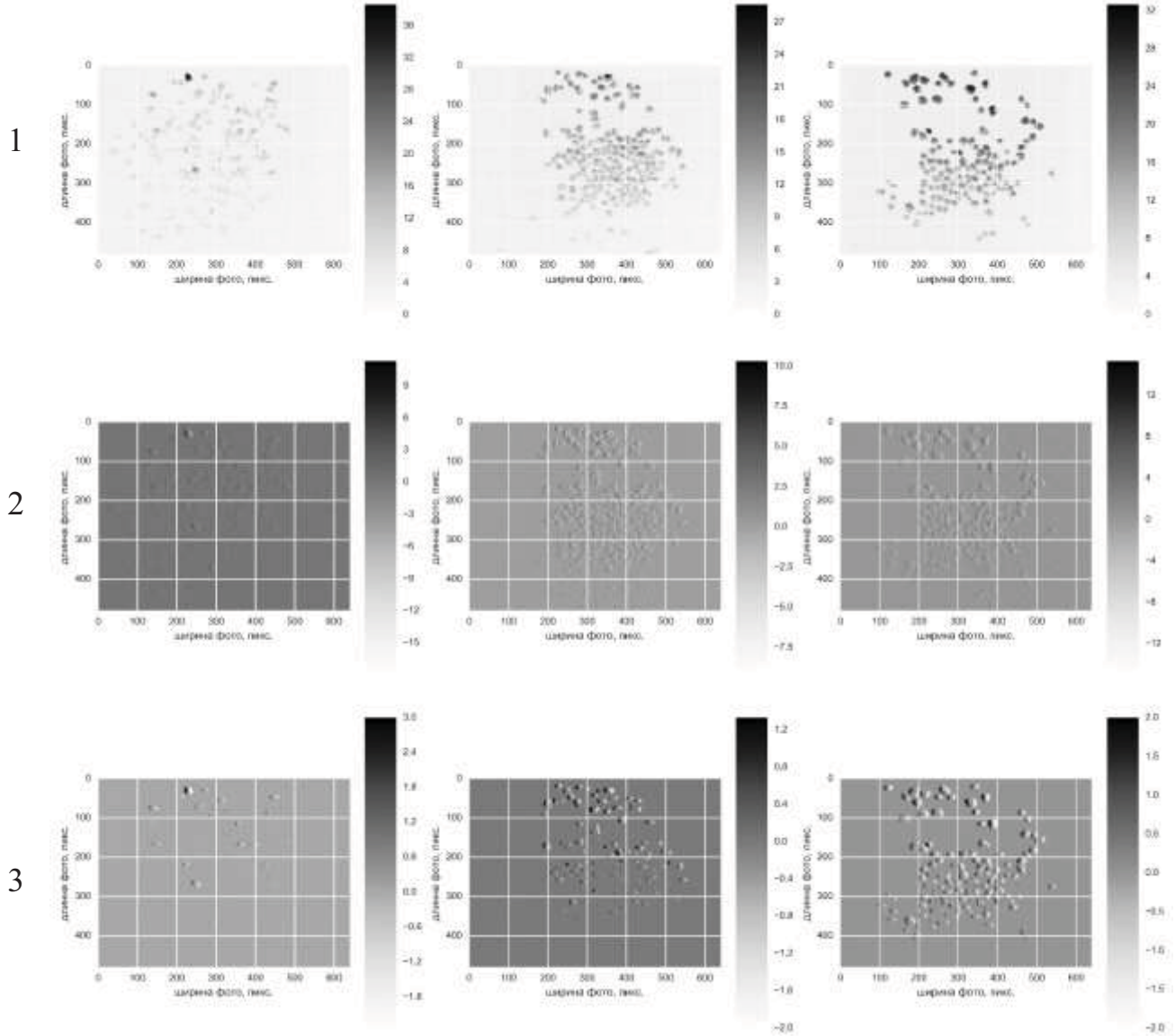


Рисунок 7.15. Пример работы алгоритма оптического регистратора: а) – гранулы NPK удобрения, промышленно обработанные к.д. 0,3 масс. %, б) – гранулы NPK удобрения, лабораторно обработанные к.д. 0,3 масс. %, в) – гранулы МАФ, лабораторно обработанные к.д. 0,3 масс. %. В строке 1 показана исходная поверхность черно-белого изображения, во 2 – поверхность после дифференцирования, в 3 – карта поверхности после сглаживания медианным фильтром.

Полученные данные о суммарной удельной площади свечения используются для сравнения качества обработки удобрений к.д. и заносятся в матрицу «объекты-признаки». По полученным данным проводится регрессионный анализ. На рисунке 7.16 приведено сравнение различных образцов минеральных удобрений по качеству обработки.

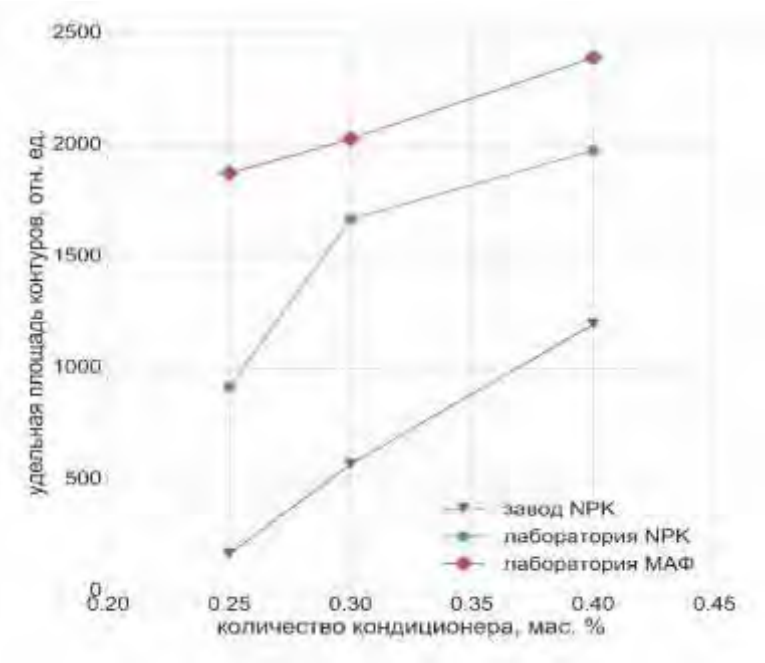


Рисунок 7.16. Сравнение минеральных удобрений по качеству обработки к.д..

По построенным графикам наблюдается разделение удобрений по качеству обработки к.д. в зависимости от суммарной площади включений. Так NPK удобрение, обработанное в промышленных условиях, показывает наихудший результат, тогда как NPK, обработанное в лабораторных условиях занимает промежуточное положение из-за неровности и неоднородности поверхности гранул, в то время как МАФ обладает наиболее правильной формой и, обработанный в лабораторных условиях, имеет наилучшее качество обработки.

7.2.3.3 Расчет проводимости разбавленных растворов удобрений

Следующим нетривиальным параметром качества выпускаемой продукции является солевой индекс (с.и.) – индекс, характеризующий солевой состав удобрения [116], данный параметр так же связывают с осмотическим давлением почвенного раствора после внесения удобрений или проводимостью 0,1 % раствора

удобрений. Солевой индекс был введен в США в 1943 году и используется в основном для оценки воздействия удобрений на растения в условиях дефицита воды. На сегодняшний день в России не существует методик оценки данного параметра.

Очевидно, что данная величина напрямую зависит от химического состава удобрений, в следствии чего нами предпринята попытка регрессионной оценки с использованием разработанного рентгенооптического комплекса. На первом этапе была составлена карта линейных корреляций солевого индекса с выделенными признаками (рисунок 7.17).

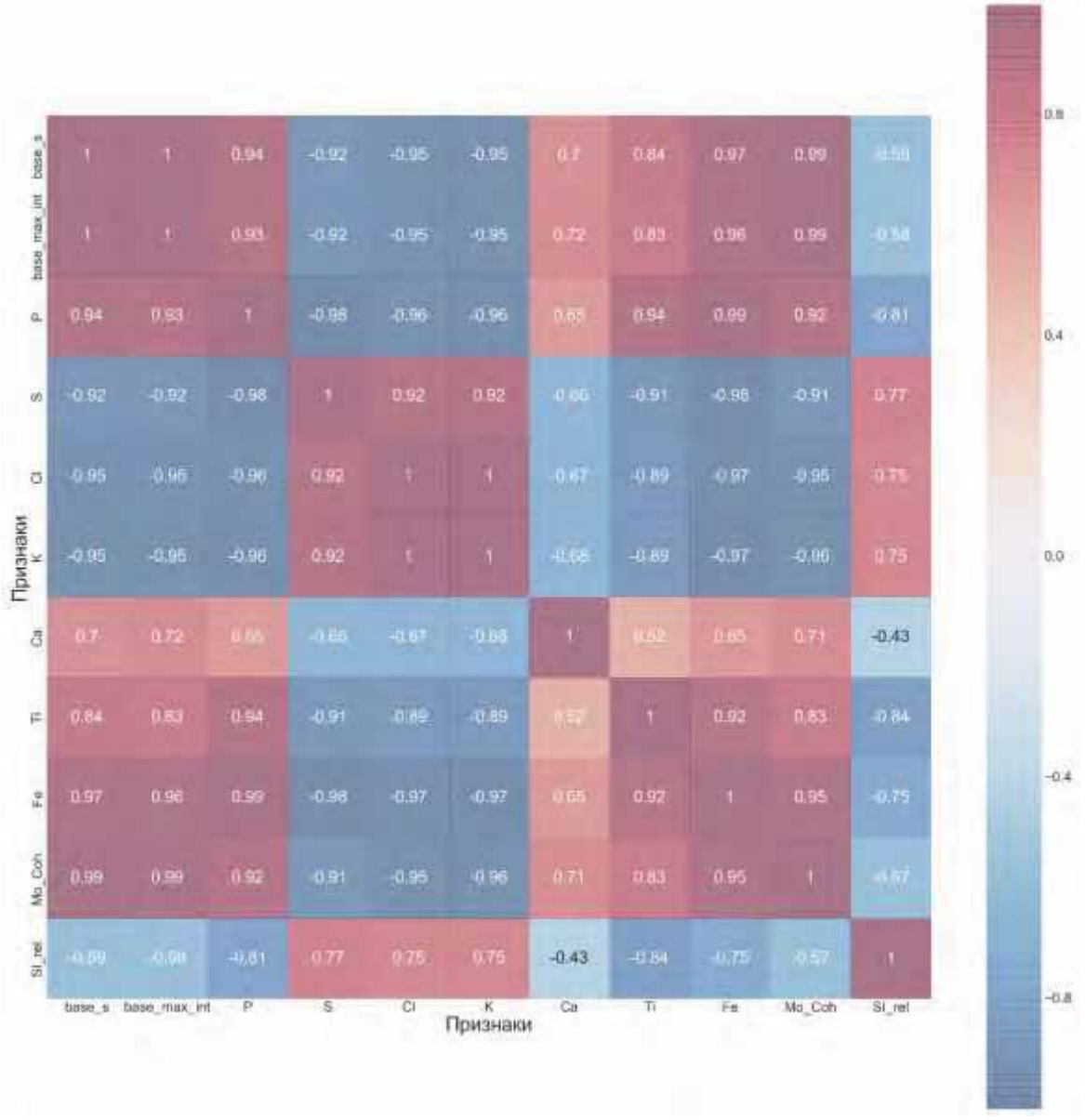


Рисунок 7.17. Карта линейных корреляций химического состава исследованных удобрений и экспериментального с.и.

Можно обратить внимание, что наиболее значимыми для предсказания с.и. являются интенсивности P, S и Ti элементов, что скорее-всего объясняется их высоким зарядом. Однако концентрация титана в удобрениях достаточно мала (менее 1 масс. %), а значит модель можно упростить. Построенное регрессионное уравнение с использованием наиболее информативных признаков имеет вид:

$$I_c = 1110.3 \cdot 1 - 0.36 \cdot I_P - 0.14 \cdot I_S + 0.007 \cdot S_{base}$$

где: I_c – исправленный аналитический сигнал, I_P – интенсивность линии фосфора, I_S – интенсивность линии серы, S_{base} – площадь рассеянного излучения.

График зависимости исправленного значения химического состава удобрений от СИ приведен на рисунке 7.18.

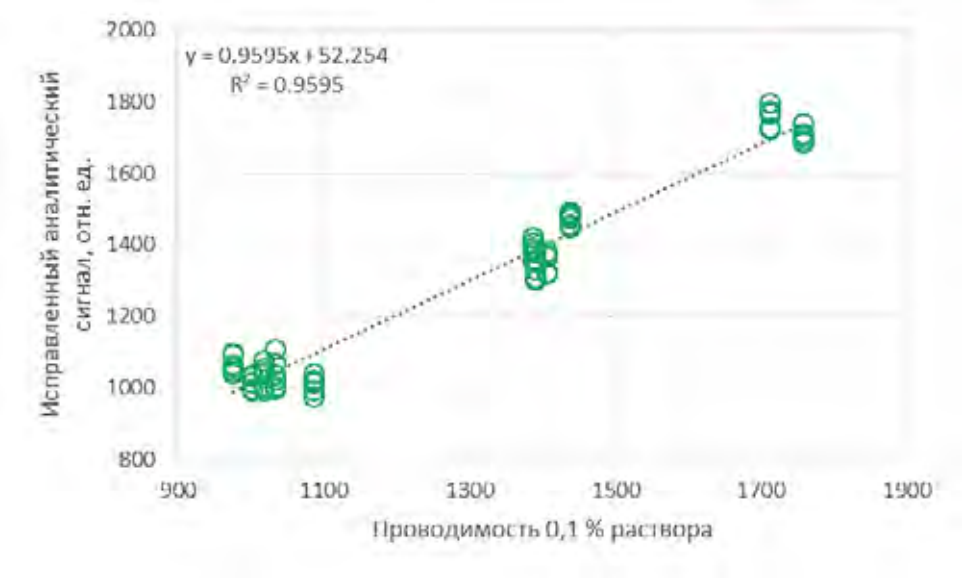


Рисунок 7.18. Зависимость солевого индекса, измеренного по проводимости 0,1 % растворов различных удобрений от их рентгенофлуоресцентных спектров.

Коэффициент корреляции (доля объясненной дисперсии) между параметрами спектра и проводимостью разбавленных растворов составляет 95,9% что является хорошим показателем для анализа на физические свойства сложных химических систем. Таким образом получена строгая зависимость солевого индекса от химического состава удобрений и предложена методика его расчета.

8 Выводы

Основные результаты диссертационной работы заключаются в следующем:

1. Выделены значимые химические и физические параметры для эффективного учета сложной матрицы исследуемых объектов, такие как: аномалии на карте поверхности (оптический регистратор), интенсивности линий химических элементов, площадь фона и интенсивность когерентного и некогерентного рассеяния рентгеновской трубки (рентгенофлуоресцентный анализ).
2. Доказана возможность определения содержания азота в сложных фосфорсодержащих удобрениях по косвенным признакам.
3. Обоснована и разработана экспрессная и мало затратная система получения информации об объектах анализа с оптимизацией стадий пробоподготовки.
4. Разработан автоматический алгоритм выделения и расчета информативных признаков при контроле качества производимой продукции.
5. Создан прототип единой аналитической базы исследуемых объектов с возможностью использования методов анализа больших данных.
6. Доказана возможность решения расширенного спектра задач разработанным в настоящей работе аппаратно-программным комплексом по сравнению с классическим методом РФА.
7. Показана возможность выявления значимых признаков объектов анализа с использованием рентгенофлуоресцентно-оптического анализатора при отсутствии явных аналитических решений поставленной задачи.
8. Разработано и внедрено программное обеспечение, обеспечивающее удобную и простую реализацию всех алгоритмов, приведенных в данной работе.

9 Список литературы

1. Abouzeid A.M. Physical and thermal treatment of phosphate ores — An overview // Int. J. Miner. Process. 2008. Vol. 85. P. 59–84.
2. Cheremisinoff N.P., Rosenfeld P. Industry and Products // Handbook of Pollution Prevention and Cleaner Production Vol. 3: Best Practices in the Agrochemical Industry. Elsevier, 2010. P. 320.
3. Cheremisinoff N.P., Rosenfeld P. Fertilizers // Handbook of Pollution Prevention and Cleaner Production Vol. 3: Best Practices in the Agrochemical Industry. Elsevier, 2010. P. 320.
4. Schorr M. Phosphate ore processing for phosphoric acid production: classical and novel technology // Miner. Process. Extr. Metall. 2010. Vol. 119, № 3. P. 125–129.
5. Sis H., Chander S. Reagents used in the flotation of phosphate ores: a critical review // Miner. Eng. 2003. Vol. 16. P. 577–585.
6. Эвенчик С.Д. и др. Технология фосфорных и комплексных удобрений. Москва: Химия, 1987. 464 p.
7. Бабкин В.В., Бродский А.А. Фосфорные удобрения России. Москва: ТОО “Агрохим-принт,” 1995. 464 p.
8. Mazzilli B. и др. Radiochemical characterization of Brazilian phosphogypsum // J. Environ. Radioact. 2000. Vol. 49. P. 113–122.
9. Pérez-lópez R. и др. Applied Geochemistry Dynamics of contaminants in phosphogypsum of the fertilizer industry of Huelva (SW Spain): From phosphate rock ore to the environment // Appl. Geochemistry. Elsevier Ltd, 2010. Vol. 25, № 5. P. 705–715.
10. Saueia C.H., Mazzilli B.P., Fávaro D.I.T. Natural radioactivity in phosphate rock , phosphogypsum and phosphate fertilizers in Brazil // J. Radioanal. Nucl. Chem. 2005. Vol. 264, № 2. P. 445–448.
11. Vito L., Cristal D. De, Dipo M. Fertilizer Characterization : Isotopic Data (N, S, O, C, and Sr) // Environ. Sci. Technol. 2004. Vol. 38, № 12. P. 3254–3262.
12. Yu J. и др. Flotation collophane from high-iron phosphate ore by using sodium ligninsulfonate as depressant // Separation Science and Technology. 2017. Vol. 52, № 3. 557-566 p.
13. Walan P. и др. Resources , Conservation and Recycling Phosphate rock production and depletion: Regional disaggregated modeling and global implications // "Resources, Conserv. Recycl. Elsevier B.V., 2014. Vol. 93. P. 178–187.
14. Макаренко М.В., Чмель С.Ю. Модернизация промышленности минеральных удобрений // Экономический журнал. Общество с ограниченной ответственностью“ Издательство Ипполитова,” 2014. Vol. 1, № 33. P. 92–103.

15. Sun K. и др. Application and Mechanism of Anionic Collector Sodium Dodecyl Sulfate (SDS) in Phosphate Beneficiation // Minerals. 2017. Vol. 7, № 2. P. 29.
16. Schorr M., Valdez B. The phosphoric acid industry: Equipment, materials, and corrosion // Corros. Rev. 2016. Vol. 34, № 1–2. P. 85–102.
17. Al-fariss T.F. и др. Investigating sodium sulphate as a phosphate depressant in acidic media // Sep. Purif. Technol. Elsevier B.V., 2014. Vol. 124. P. 163–169.
18. Al-Fariss T.F. и др. Low Solubility of Calcined Phosphate: Surface Area Reduction or Chemical Composition Change? // Part. Sci. Technol. 2014. Vol. 32, № 1. P. 80–85.
19. Boulos T.R. и др. A modification in the flotation process of a calcareous – siliceous phosphorite that might improve the process economics // Miner. Eng. Elsevier Ltd, 2014. Vol. 69. P. 97–101.
20. Frikha N., Hmercha A., Gabsi S. Modelling of a solid dissolution in liquid with chemical reaction: Application to the attack reaction of phosphate by sulphuric acid // Can. J. Chem. Eng. 2014. Vol. 92, № 10. P. 1829–1838.
21. Melike S., Özer A.K., Gülaboglu M.S. Investigation of the changes of P2O5 content of phosphate rock during simultaneous calcination / sulfation // Powder Technol. 2011. Vol. 211. P. 72–76.
22. Петропавловский И.А. и др. Оценка возможности обогащения и химической переработки некондиционного фосфатного сырья на основе исследования химического и минералогического состава // Химическая промышленность сегодня. Общество с ограниченной ответственностью“ Химпром сегодня,” 2012. № 4. P. 5–8.
23. ТУ 2111-040-00203938-98 Концентрат апатитовый загрубленного помола. 1998. P. 26.
24. Дормешкин О.Б., Черчес Г.Х., Гаврилюк А.Н. Влияние видов фосфатного сырья на технологический процесс производства экстракционной фосфорной кислоты и комплексных удобрений // Труды БГТУ. Серия 3 Химия и технология неорганических веществ. 2013. Vol. 3, № 1. P. 71–76.
25. ТУ 2186-689-00209438-09 Удобрения азотно-фосфорно-калийное. 2009. P. 38.
26. ТУ 2186-687-00209438-08. Удобрение азотно-фосфорное серосодержащее. 2010. P. 30.
27. ТУ 2186-00209438-2014. Диаммонийфосфат удобрительный. 2015. P. 24.
28. ГОСТ Р 51520-99. Удобрения минеральные. Общие технические условия. 2000. P. 9.
29. Федотов П.С., Норов А.М., Петропавловский И.А. Новая гибкая технология получения гранулированных сложных серосодержащих фосфорно-калийных удобрений // Актуальные направления научных исследований от теории к практике. 2015. № 2. P. 137–140.
30. Sivarajah U. и др. Critical analysis of Big Data challenges and analytical methods // J. Bus. Res. The Authors, 2017. Vol. 70. P. 263–286.

31. Offroy M., Duponchel L. Topological data analysis: a promising big data exploration tool in biology, analytical chemistry and physical chemistry // Anal. Chim. Acta. Elsevier Ltd, 2016.
32. Hasikova J. и др. On-Line XRF analysis of phosphate materials at various stages of processing // Procedia Eng. Elsevier B.V., 2014. Vol. 83. P. 455–461.
33. Sokolov A.D. и др. On-line analysis of chrome-iron ores on a conveyor belt using x-ray fluorescence analysis // X-Ray Spectrom. 2005. Vol. 34, № 5. P. 456–459.
34. Remes A., Saloheimo K., Jämsä-Jounela S.L. Effect of speed and accuracy of on-line elemental analysis on flotation control performance // Miner. Eng. 2007. Vol. 20, № 11. P. 1055–1066.
35. Лобоцкий Ю.Г., Хмара В.В. Автоматизированные системы аналитического контроля как основа управления технологическим процессом // Новое слово в науке и практике гипотезы и апробация результатов исследований. 2013. № 7. P. 146–151.
36. Cernik M., Borkovec M., Westall J.C. Regularized Least-Squares Methods for the Calculation of Discrete and Continuous Affinity Distributions for Heterogeneous Sorbents // Environ. Sci. Technol. 1995. Vol. 29, № 2. P. 413–425.
37. Cutler D.R. и др. Random Forests for Classification in Ecology // Ecology. 2007. Vol. 88, № 11. P. 2783–2792.
38. Gaspar H.A. и др. Chemical Data Visualization and Analysis with I ncremental GTM : Big Data Challenge // J Chem Inf Model. 2015. Vol. 55, № 1. P. 84–94.
39. Graaf C. De и др. Managing the Computational Chemistry Big Data Problem : The ioChem-BD Platform // J Chem Inf Model. 2015. Vol. 55. P. 95–103.
40. Jones G. и др. Application of the Bootstrap to Calibration Experiments // Anal. Chem. 1996. Vol. 68, № 5. P. 763–770.
41. Kunz M.R., Kalivas J.H., Andries E. Model updating for spectral calibration maintenance and transfer using 1-norm variants of tikhonov regularization // Anal. Chem. 2010. Vol. 82, № 9. P. 3642–3649.
42. Wehrens R., Van Der Linden W.E. Bootstrapping principal component regression models // J. Chemommetrics. 1997. Vol. 11, № September 1996. P. 157–171.
43. Svetnik V. и др. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling // J. Chem. Inf. Comput. Sci. 2003. Vol. 43, № 6. P. 1947–1958.
44. Schaik C. Van и др. Data-driven medicinal chemistry in the era of big data // Drug Discov. Today. 2014. Vol. 0, № 0. P. 1–10.
45. Palmer D. Random forest models to predict aqueous solubility // J Chem Inf Model. 2007. Vol. 47. P. 150–158.

46. Moran K. Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays To Identify Potential Toxicants // Chem. Res. Toxical. 2014. Vol. 27, № 10. P. 1643–1651.
47. Liu M. и др. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar // Sensors Actuators, B Chem. Elsevier B.V., 2013. Vol. 177. P. 970–980.
48. Bonate P.L. Approximate Confidence Intervals in Calibration Using the Bootstrap // Anal. Chem. 1993. Vol. 65. P. 1367–1372.
49. Cutler A., Stevens J.R. Random Forests for Microarrays // Methods Enzymol. 2006. Vol. 411, № 6. P. 422–432.
50. Montenegro-burke J.R. и др. Data Streaming for Metabolomics : Accelerating Data Processing and Analysis from Days to Minutes Data Streaming for Metabolomics : Accelerating Data Processing and Analysis from Days to Minutes // Anal. Chem. 2017. Vol. 89, № 2. P. 1254–1259.
51. Breiman L. Random forests // Mach. Learn. 2001. Vol. 45, № 1. P. 5–32.
52. Shih W.C., Bechtel K.L., Feld M.S. Constrained regularization: Hybrid method for multivariate calibration // Anal. Chem. 2007. Vol. 79, № 1. P. 234–239.
53. Зинин Д.С., Бушуев Н.Н., Кузнецов В.В. Рентгенофлуоресцентное определение La, Ce, Pr, Nd и Sm в промышленных осадках сульфата кальция с использованием линейного регрессионного анализа // Журнал Аналитической Химии. 2017. Vol. 72, № 3. P. 226–237.
54. Юновидов Д.В. и др. Разработка стандартного образца апатитового концентрата. Эффективный контроль однородности с помощью рентгенофлуоресцентных методов анализа. // ГИАБ. 2016. Vol. 7. P. 131–144.
55. ГОСТ 27872-88. Стандартные образцы (состава горных пород). 1989. P. 49.
56. ГОСТ 8.532-2002. Стандартные образцы состава веществ и материалов. Межлабораторная метрологическая аттестация. Содержание и порядок проведения работ. 2010. P. 13.
57. МИ 2083-90. Измерения косвенные. Определение результатов измерении и оценивание их погрешностей. 1991. P. 16.
58. ГОСТ Р ИСО 5725-2-2002. Точность (правильность и прецезионность) методов и результатов измерений. Часть 2. Основной метод определения повторяемости и воспроизводимости стандартного метода измерений. 2002. P. 62.
59. МИ 2335-2003. Внутренний контроль качества результатов количественного химического анализа. 2004. P. 67.
60. ГОСТ Р 50779.10-2000. Статистические методы. Вероятность и основы статистики.

Термины и определения. 2000. P. 46.

61. РМГ 61-2003. Показатели точности, правильности, прецизионности методик количественного химического анализа. 2007. P. 41.
62. ГОСТ 8.315-97. Стандартные образцы состава и свойств веществ и материалов. 2004. P. 26.
63. Weindorf D.C., Bakr N., Zhu Y. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications // Adv. Agron. Elsevier, 2014. Vol. 128. P. 1–45.
64. Estela J.M., Cerdà V. Flow analysis techniques for phosphorus: An overview // Talanta. 2005. Vol. 66, № 2 SPEC. ISS. P. 307–331.
65. Кристиан Г. Аналитическая химия. Москва: Бином, 2009. 623 p.
66. Золотов Ю.А., Дорохова Е.Н., Фадеева В.И. Основы аналитической химии. Москва: Высшая школа, 2004. 503 p.
67. Issahary D., Pelly I. Simultaneous multielement analysis of phosphates by x-ray fluorescence // X-Ray Spectrom. 1982. Vol. 11, № 1. P. 8–12.
68. Marvin D.C., Ives N.A. Real Time Chemical Analysis of Phosphoric - Acid Using Energy Dispersive X-Ray Fluorescence // X-Ray Spectrom. 1983. Vol. 12, № 3. P. 106–110.
69. Alexandrova A., Arpadjan S. Column solid phase extraction as preconcentration method for trace element determination in oxalic acid by atomic absorption spectrometry and inductively coupled plasma atomic emission spectrometry // Anal. Chim. Acta. 1995. Vol. 307, № 1. P. 71–77.
70. Wankova J. и др. X-Ray Fluorescence Analysis of Natural Phosphates // X-Ray Spectrom. 1982. Vol. 11, № 3. P. 109–111.
71. Chaerun S.K. Tempeh Waste as a Natural, Economical Carbon and Nutrient Source: ED-XRF and NCS Study // HAYATI J. Biosci. 2009. Vol. 16, № 3. P. 120–122.
72. Alcalde-Molina M., Ruiz-Jiménez J., Luque de Castro M.D. Automated determination of mercury and arsenic in extracts from ancient papers by integration of solid-phase extraction and energy dispersive X-ray fluorescence detection using a lab-on-valve system // Anal. Chim. Acta. 2009. Vol. 652, № 1–2. P. 148–153.
73. Butcher D.J. Advances in Inductively Coupled Plasma Optical Emission Spectrometry for Environmental Analysis // Instrum. Sci. Technol. 2010. Vol. 38, № 6. P. 458–469.
74. Weindorf D.C. и др. Use of portable X-ray fluorescence spectrometry for environmental quality assessment of peri-urban agriculture // Environ. Monit. Assess. 2012. Vol. 184, № 1. P. 217–227.
75. Chauhan P., Chauhan R.P. Elemental analysis of fertilizers using X-ray fluorescence and their impact on alpha radioactivity of plants // J. Radioanal. Nucl. Chem. 2013. Vol. 295, № 2. P. 1097–1105.

76. De Oliveira Souza S. и др. Simultaneous determination of macronutrients, micronutrients and trace elements in mineral fertilizers by inductively coupled plasma optical emission spectrometry // Spectrochim. Acta - Part B At. Spectrosc. Elsevier B.V., 2014. Vol. 96. P. 1–7.
77. De La Calle I. и др. Fast method for multielemental analysis of plants and discrimination according to the anatomical part by total reflection X-ray fluorescence spectrometry // Food Chem. Elsevier Ltd, 2013. Vol. 138, № 1. P. 234–241.
78. West M. и др. Atomic spectrometry update–X-ray fluorescence spectrometry // J. Anal. At. Spectrom. 2012. Vol. 27. P. 1603–1644.
79. Rui Y. kui, Hao J., Rui F. Determination of seven plant nutritional elements in potassium dihydrogen phosphate fertilizer from northeastern China // J. Saudi Chem. Soc. King Saud University, 2012. Vol. 16, № 1. P. 89–90.
80. Krachler M. Environmental applications of single collector high resolution ICP-MS // J. Environ. Monit. 2007. Vol. 9, № 8. P. 790–804.
81. Safi M.J. и др. Chemical analysis of phosphate rock using different methodsadvantages and disadvantages // X-Ray Spectrom. 2006. Vol. 35, № 3. P. 154–158.
82. Муханова А.А. и др. Рентгенофлуоресцентный анализ водно-органических технологических растворов // Заводская лаборатория. Диагностика материалов. 2006. Vol. 72, № 10. P. 18–22.
83. Al-Shawi A.W., Dahl R. Determination of total chromium in phosphate rocks by ion chromatography // J. Chromatogr. A. 1999. Vol. 850, № 1–2. P. 137–141.
84. Icardo M.C. и др. Flow spectrophotometric determination of ammonium ion // Anal. Chim. Acta. 1999. Vol. 398, № 2–3. P. 311–318.
85. Burns D.T. и др. Flow-injection spectrophotometric determination of phosphate using Crystal Violet // Anal. Chim. Acta. 1991. Vol. 254, № 1–2. P. 197–200.
86. Кузьмина Т.Г. и др. О погрешности пробоподготовки при прессовании излучателей для рентгенофлуоресцентного анализа // Журнал Аналитической Химии. 2017. Vol. 72, № 3. P. 218–226.
87. Кривоносов В.А. и др. Математическая модель процесса экстракции и фильтрации производства фосфорной кислоты ООО “Балаковские минеральные удобрения” // Автоматизация в промышленности. 2013. Vol. 15. P. 24–29.
88. https://www.scopus.com/ [Electronic resource]. 2017.
89. Shihe L., Shengquan L., Yuejun Z. Spectrophotometric Determination of the Total Amount of Rare-Earth Elements in Agricultural Samples with p-Chloro-chlorophosphonazo // Talanta. 1992. Vol. 39, № 8. P. 987–991.

90. Sharma A. и др. Characterizing soils via portable X-ray fluorescence spectrometer: 4. Cation exchange capacity (CEC) // Geoderma. Elsevier B.V., 2015. Vol. 239. P. 130–134.
91. Chauhan P., Chauhan R.P., Gupta M. Estimation of naturally occurring radionuclides in fertilizers using gamma spectrometry and elemental analysis by XRF and XRD techniques // Microchem. J. Elsevier B.V., 2013. Vol. 106. P. 73–78.
92. BS ISO 12926:2012. Aluminium fluoride for industrial use - Determination of trace elements - Wavelength dispersive X-ray fluorescence spectrometri method using pressed powder tablets. 2012. P. 22.
93. Schroeder B. и др. Analysis of geologic materials using an automated x-ray fluorescence system // X-Ray Spectrom. 1980. Vol. 9, № 4. P. 198–205.
94. Menegario A.A. и др. On-line preconcentration flow system for multi-elemental analysis by total reflection X-ray fluorescence spectrometry // Spectrochim. Acta Part B At. Spectrosc. 2003. Vol. 58, № 3. P. 543–549.
95. Docenko D. и др. On-line measurement of uranium in ores using XRF analyzer P2 Con-X. 2005. 29 p.
96. Ермолинская В. и др. Автоматизированная система измерения концентраций металлов в технологических растворах на базе портативного РФА-анализатора “X-SPEC” // Аналитика. 2015. Vol. 2. P. 110–115.
97. Gemelli M., D’Orazio M., Folco L. Chemical analysis of iron meteorites using a hand-held X-ray fluorescence spectrometer // Geostand. Geoanalytical Res. 2015. Vol. 39, № 1. P. 55–69.
98. Marks M.A.W. и др. The volatile inventory (F, Cl, Br, S, C) of magmatic apatite: An integrated analytical approach // Chem. Geol. Elsevier B.V., 2012. Vol. 291. P. 241–255.
99. Shand C.A., Wendler R. Portable X-ray fluorescence analysis of mineral and organic soils and the influence of organic matter // J. Geochemical Explor. 2014. Vol. 143. P. 31–42.
100. Angeyo K.H. и др. Optimization of X-ray fluorescence elemental analysis: An example from Kenya // Appl. Radiat. Isot. 1998. Vol. 49, № 7. P. 885–891.
101. Hu W. и др. Metals analysis of agricultural soils via portable x-ray fluorescence spectrometry // Bull. Environ. Contam. Toxicol. 2014. Vol. 92, № 4. P. 420–426.
102. Pukhovski A. V. X-ray fluorescence analysis in the Russian State Agrochemical Service: An overview // X-Ray Spectrom. 2002. Vol. 31, № 3. P. 225–234.
103. Towett E.K., Shepherd K.D., Cadisch G. Quantification of total element concentrations in soils using total X-ray fluorescence spectroscopy (TXRF) // Sci. Total Environ. 2013. Vol. 463–464. P. 374–388.
104. Plebow A. X-ray-induced alteration of specimens as crucial obstacle in XRF spectrometry of

fluorine in rocks and soils // X-Ray Spectrom. 2013. Vol. 42, № 1. P. 19–32.

105. Gullayanon R. A calibration methodology for energy dispersive X-Ray fluorescence measurements based upon synthetically generated reference spectra. 2011. 223 p.
106. Stockmann U. и др. Utilizing portable X-ray fluorescence spectrometry for in-field investigation of pedogenesis // Catena. 2016. Vol. 139. P. 220–231.
107. Осколок К.В., Моногарова О.В., Гармай А.В. Количественный рентгенофлуоресцентный анализ многоэлементных объектов сложной формы без использования образцов сравнения // Вестник Московского университета. Серия 2. Химия. 2014. Vol. 55, № 1. P. 10–14.
108. Bueno Guerra M.B. и др. Comparison of analytical performance of benchtop and handheld energy dispersive X-ray fluorescence systems for the direct analysis of plant materials // J. Anal. At. Spectrom. 2014. Vol. 29, № 9. P. 1667.
109. Angeyo K.H. и др. Feasibility for direct rapid energy dispersive X-ray fluorescence (EDXRF) and scattering analysis of complex matrix liquids by partial least squares // Appl. Radiat. Isot. Elsevier, 2012. Vol. 70, № 11. P. 2596–2601.
110. Marguí E. и др. Analytical possibilities of different X-ray fluorescence systems for determination of trace elements in aqueous samples pre-concentrated with carbon nanotubes // Spectrochim. Acta - Part B At. Spectrosc. Elsevier B.V., 2013. Vol. 88. P. 192–197.
111. Bosco G.L. Development and application of portable, hand-held X-ray fluorescence spectrometers // TrAC - Trends Anal. Chem. 2013. Vol. 45. P. 121–134.
112. Rousseau R.M., Willis J.P., Duncan A.R. Practical XRF Calibration Procedures for Major and Trace Elements // X-Ray Spectrom. 1996. Vol. 25, № January. P. 179–189.
113. Rousseau R.M. A Comprehensive Alpha Coefficient Algorithm (a Second Version) // X-Ray Spectrom. 1987. Vol. 16, № February 1985. P. 103–108.
114. Rousseau R.M. Corrections for matrix effects in X-ray fluorescence analysis-A tutorial // Spectrochim. Acta - Part B At. Spectrosc. 2006. Vol. 61, № 7. P. 759–777.
115. Strugailo V. Review of filtration and segmentation methods for digital images // Sci. Educ. Bauman MSTU. 2012. Vol. 12, № 5. P. 270–281.
116. Юновидов Д.В. и др. Солевой индекс // Роль аналитических служб в обеспечении качества минеральных удобрений и серной кислоты. 2015. 23-35 p.

Приложение А

В данном приложении приведена программная реализация основных функций и алгоритмов, использованных в работе (на языке Python 2.7). Листинг составлен в среде разработке "jupyter" и не является исполняемым. Данные и код, использованные в работе можно запросить у автора по адресу Dm.Yunovidov@gmail.com.

```
# Импортирование модулей
import numpy as np
import matplotlib as mpl
from matplotlib import rc
import math
import pandas as pd
import scipy
import seaborn as sns
from sklearn.cross_validation import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report, accuracy_score, f1_score
from sklearn.metrics import precision_score, recall_score, mean_squared_error
from sklearn import cross_validation, datasets, linear_model, metrics, grid_search
from sklearn.preprocessing import OneHotEncoder
from sklearn import random_projection
from sklearn.decomposition import PCA
from sklearn import manifold
from sklearn.cluster import KMeans
from sklearn.neighbors import KNeighborsClassifier
sns.set_style("whitegrid")
sns.set_palette('Accent')
rc('font', family='Arial')
% matplotlib inline
print 'Import Ready'

# Функции для классификации
def classif(all_data, all_labels, tdata=''):
    '''
    Function for optimmize classification and calculate main quality metrics.
    Return:
        outputs with data
    Parameters:
        train_data
        train_labels
        test_data - data for test
        test_labels - labels for test
        tdata
    '''
    # Split to train and test
    train_data, test_data, train_labels, test_labels =
cross_validation.train_test_split(all_data, all_labels,

test_size=0.2,train_size=0.8,stratify=all_labels)
    print 'Test train-test split: ',
    print Counter(all_labels),
    print Counter(train_labels),
    print Counter(test_labels)
```

```
    cv = cross_validation.StratifiedShuffleSplit(train_labels, n_iter = 10,
test_size = 0.3, random_state=0)
    cv_metricks = cross_validation.StratifiedShuffleSplit(all_labels, n_iter = 10,
test_size = 0.3, random_state=0)
    metrics_all = ['accuracy', 'precision_macro', 'f1_macro', 'recall_macro']
    metriks_names = ['accuracy', 'precision_macro', 'f1_macro', 'recall_macro']

    # 1 - Liner classifire with SGD
    print '\tSGD Liner classifire:'
    log_regressor = linear_model.SGDClassifier(penalty='none', shuffle=True,
random_state=0)
    parameters_grid = {
        'loss': ['hinge', 'log','squared_loss', 'modified_huber'],
        'fit_intercept': [True, False],
        'n_iter' : np.linspace(1000, 15000, 5, dtype=int),
    }
    grid_cv = grid_search.GridSearchCV(log_regressor, parameters_grid, scoring =
'f1_macro', cv = cv)
    grid_cv.fit(train_data, train_labels)
    print '  Best with grid search: '
    print '  estimator: ', grid_cv.best_estimator_
    print '  score: ', grid_cv.best_score_
    print '  parameters: ', grid_cv.best_params_
    print '  Features importamce: '
    a = train_data.columns
    imp = np.mean(np.abs(grid_cv.best_estimator_.coef_),0)
    b = imp / np.sum(imp) * 100.
    importances = pd.DataFrame(zip(a, b))
    importances.columns = ['feature name', 'importance']
    print '  ', importances.sort_values(by='importance', ascending=False)
    print "  Detailed classification report:"
    y_true, y_pred = test_labels, grid_cv.best_estimator_.predict(test_data)
    print '  ', classification_report(y_true, y_pred)
    for i in xrange(len(metrics_all)):
        scor = metrics_all[i]
        scoring = cross_validation.cross_val_score(grid_cv.best_estimator_,
all_data, all_labels, scoring = scor, cv = cv_metricks)
        print '  Best SGD Lin ' + \
            metriks_names[i] + \
            ' mean:{}, max:{}, min:{}, std:{}'.format(scoring.mean(),
scoring.max(), scoring.min(), scoring.std())

    # 2 - Ridge (L2) classifire
    print
    ridge_classifier = linear_model.SGDClassifier(penalty='l2',shuffle=True,
random_state=0)
    print '\tRidge Classifire (SGD with L2):'
    parameters_grid = {
        'loss': ['hinge', 'log','squared_loss', 'modified_huber'],
        'fit_intercept': [True, False],
        'n_iter': np.linspace(1000,15000, 5, dtype=int),
        'alpha': np.linspace(0.0001, 1., num = 10)
    }
    grid_cv = grid_search.GridSearchCV(ridge_classifier, parameters_grid, scoring
= 'f1_macro', cv = cv)
    grid_cv.fit(train_data, train_labels)
    print '  Best with grid search: '
    print '  estimator: ', grid_cv.best_estimator_
    print '  score: ', grid_cv.best_score_
    print '  parameters: ', grid_cv.best_params_
    print '  Features importamce: '
    a = train_data.columns
    imp = np.mean(np.abs(grid_cv.best_estimator_.coef_),0)
```

```
    b = imp / np.sum(imp) * 100.
    importances = pd.DataFrame(zip(a, b))
    importances.columns = ['feature name', 'importance']
    print '  ', importances.sort_values(by='importance', ascending=False)
    print "  Detailed classification report:"
    y_true, y_pred = test_labels, grid_cv.best_estimator_.predict(test_data)
    print '  ', classification_report(y_true, y_pred)
    for i in xrange(len(metrics_all)):
        scor = metrics_all[i]
        scoring = cross_validation.cross_val_score(grid_cv.best_estimator_,
all_data, all_labels, scoring = scor, cv = cv_metricks)
        print '  Best Ridge ' + \
            metriks_names[i] + \
            ' mean:{}, max:{}, min:{}, std:{}'.format(scoring.mean(),
scoring.max(), scoring.min(), scoring.std())

    # 3 - Lasso (L1) classifire
    print
    print '\tLasso Classifire (SGD with L1):'
    lasso_classifire = linear_model.SGDClassifier(penalty='l1',shuffle=True,
random_state=0)
    parameters_grid = {
        'loss': ['hinge', 'log','squared_loss', 'modified_huber'],
        'fit_intercept': [True, False],
        'n_iter': np.linspace(1000,15000, 5, dtype=int),
        'alpha': np.linspace(0.0001, 1., num = 10)
    }
    grid_cv = grid_search.GridSearchCV(lasso_classifire, parameters_grid, scoring
= 'f1_macro', cv = cv)
    grid_cv.fit(train_data, train_labels)
    print '  Best with grid search: '
    print '  estimator: ', grid_cv.best_estimator_
    print '  score: ', grid_cv.best_score_
    print '  parameters: ', grid_cv.best_params_
    print '  Features importamce: '
    a = train_data.columns
    imp = np.mean(np.abs(grid_cv.best_estimator_.coef_),0)
    b = imp / np.sum(imp) * 100.
    importances = pd.DataFrame(zip(a, b))
    importances.columns = ['feature name', 'importance']
    print '  ', importances.sort_values(by='importance', ascending=False)
    print "  Detailed classification report:"
    y_true, y_pred = test_labels, grid_cv.best_estimator_.predict(test_data)
    print '  ', classification_report(y_true, y_pred)

    for i in xrange(len(metrics_all)):
        scor = metrics_all[i]
        scoring = cross_validation.cross_val_score(grid_cv.best_estimator_,
all_data, all_labels, scoring = scor, cv = cv_metricks)
        print '  Best L1 ' + \
            metriks_names[i] + \
            ' mean:{}, max:{}, min:{}, std:{}'.format(scoring.mean(),
scoring.max(), scoring.min(), scoring.std())

    # 4 - Random Forest
    print
    rf_classifier = RandomForestClassifier(random_state=0)
    print '\tRandom Forest:'
    parameters_grid = {
        'n_estimators' : range(2, 100, 20),
        'max_features' : ['auto', 'sqrt', 'log2', None],
        'max_depth': [None] + range(2,13,5),
        'bootstrap' : [False, True],
```

```
        'class_weight': ['balanced', None]
    }
    grid_cv = grid_search.GridSearchCV(rf_classifier, parameters_grid, scoring =
'f1_macro', cv = cv)
    grid_cv.fit(train_data, train_labels)
    print '  Best with grid search: '
    print '  estimator: ', grid_cv.best_estimator_
    print '  score: ', grid_cv.best_score_
    print '  parameters: ', grid_cv.best_params_
    print '  Features importamce: '
    importances = pd.DataFrame(zip(train_data.columns,
grid_cv.best_estimator_.feature_importances_ * 100.))
    importances.columns = ['feature name', 'importance']
    print '  ', importances.sort_values(by='importance', ascending=False)
    print "  Detailed classification report:"
    y_true, y_pred = test_labels, grid_cv.best_estimator_.predict(test_data)
    print '  ', classification_report(y_true, y_pred)
    for i in xrange(len(metrics_all)):
        scor = metrics_all[i]
        scoring = cross_validation.cross_val_score(grid_cv.best_estimator_,
all_data, all_labels, scoring = scor, cv = cv_metricks)
        print '  Best RF ' + \
            metriks_names[i] + \
            ' mean:{}, max:{}, min:{}, std:{}'.format(scoring.mean(),
scoring.max(), scoring.min(), scoring.std())

    # 5 - Baess
    print
    b_cl = MultinomialNB()
    b_cl.fit(train_data.abs(), train_labels)
    print '\tBaess: '
    try:
        for i in xrange(len(metrics_all)):
            scor = metrics_all[i]
            scoring = cross_validation.cross_val_score(b_cl, all_data, all_labels,
scoring = scor, cv = cv_metricks)
            print '  Baess ' + \
                metriks_names[i] + \
                ' mean:{}, max:{}, min:{}, std:{}'.format(scoring.mean(),
scoring.max(), scoring.min(), scoring.std())
    except:
        b_scoring = cross_validation.cross_val_score(b_cl, all_data.abs(),
all_labels, scoring = 'accuracy', cv = cv_metricks)
        print '  Baess accuracy mean:{}, max:{}, min:{},
std:{}'.format(b_scoring.mean(), b_scoring.max(), b_scoring.min(),
b_scoring.std())

# Функции для регрессии
def regr(all_x, all_y, name):
    '''
    Function for optimmize regression and calculate main quality metrics: absolute
error, square_error and r2.
    Return:
        outputs with data
    Parameters:
        train_data (x)
        train_labels (y)
        test_data - data for test (x_test)
        test_labels - labels for test  (y_test)
        name of target data
    '''
```

```
    x, x_test, y, y_test = cross_validation.train_test_split(all_x,all_y,
test_size=0.2, train_size=0.8, stratify=all_y)
    print 'Test train-test split: ',
    print Counter(all_y),
    print Counter(y),
    print Counter(y_test)
    cv = cross_validation.StratifiedShuffleSplit(y, n_iter = 10, test_size = 0.3,
random_state=0)
    cv_metricks = cross_validation.StratifiedShuffleSplit(all_y, n_iter = 10,
test_size = 0.3, random_state=0)
    metrics_all = ['mean_absolute_error', 'mean_squared_error', 'r2']

    print '\tLinear: '
    linear_regressor = linear_model.LinearRegression()
    linear_regressor.fit(x, y)
    print '  Features importance: '
    a = x.columns
    imp = np.abs(linear_regressor.coef_)
    b = imp / np.sum(imp) * 100.
    importances = pd.DataFrame(zip(a, b))
    importances.columns = ['feature name', 'importance']
    print importances.sort_values(by='importance', ascending=False)
    for score in metrics_all:
        linear_scoring = cross_validation.cross_val_score(linear_regressor, all_x,
all_y, scoring = score, cv = cv_metricks)
        print '  Linear ' + score + ' mean: {}, std: {}, rel std:
{}'.format(linear_scoring.mean(),linear_scoring.std(),linear_scoring.std())

    print
    print '\tLasso (L1): '
    lasso_regressor = linear_model.Lasso()
    parameters_grid = {
        'fit_intercept': [True, False],
        'normalize': [True, False],
        'max_iter': range(2, 20, 2),
        'alpha': np.linspace(0.0001, 2., num = 10)
    }
    grid_cv_lasso = grid_search.GridSearchCV(lasso_regressor, parameters_grid,
scoring = 'r2', cv = cv)
    grid_cv_lasso.fit(x, y)
    print '  Best with grid search: '
    print '  estimator: ', grid_cv_lasso.best_estimator_
    print '  score: ', grid_cv_lasso.best_score_
    print '  parameters: ', grid_cv_lasso.best_params_
    print '  Features importamce: '
    a = x.columns
    imp = np.abs(grid_cv_lasso.best_estimator_.coef_)
    b = imp / np.sum(imp) * 100.
    importances = pd.DataFrame(zip(a, b))
    importances.columns = ['feature name', 'importance']
    print '\t', importances.sort_values(by='importance', ascending=False)

    for score in metrics_all:
        lasso_scoring =
cross_validation.cross_val_score(grid_cv_lasso.best_estimator_, all_x, all_y,
scoring = score, cv = cv_metricks)
        print '  Best Lasso ' + score + ' mean: {}, std: {}, rel std:
{}'.format(lasso_scoring.mean(), lasso_scoring.std(), lasso_scoring.std())

    print
    print '\tRidge (L2): '
    ridge_regressor = linear_model.Ridge()
    parameters_grid = {
```

```
        'fit_intercept': [True, False],
        'normalize': [True, False],
        'solver' : ['auto', 'svd', 'cholesky', 'lsqr', 'sparse_cg', 'sag'],
        'max_iter': range(2, 20, 2),
        'alpha': np.linspace(0.0001, 2., num = 10)
    }
    grid_cv_ridge = grid_search.GridSearchCV(ridge_regressor, parameters_grid,
scoring = 'r2', cv = cv)
    grid_cv_ridge.fit(x, y)
    print '  Best with grid search: '
    print '  estimator: ', grid_cv_ridge.best_estimator_
    print '  score: ', grid_cv_ridge.best_score_
    print '  parameters: ', grid_cv_ridge.best_params_
    print '  Features importance: '
    a = x.columns
    imp = np.abs(grid_cv_ridge.best_estimator_.coef_)
    b = imp / np.sum(imp) * 100.
    importances = pd.DataFrame(zip(a, b))
    importances.columns = ['feature name', 'importance']
    print '\t', importances.sort_values(by='importance', ascending=False)

    for score in metrics_all:
        ridge_scoring =
cross_validation.cross_val_score(grid_cv_ridge.best_estimator_, all_x, all_y,
scoring = score, cv = cv_metricks)
        print '  Best Ridge ' + score + ' mean: {}, std: {}, rel std:
{}'.format(ridge_scoring.mean(), ridge_scoring.std(), ridge_scoring.std())

    print
    pred_lin = linear_regressor.predict(all_x)
    pred_lasso = grid_cv_lasso.best_estimator_.predict(all_x)
    pred_ridge = grid_cv_ridge.best_estimator_.predict(all_x)
    sns.set(font_scale=2, style='whitegrid', palette='Accent')
    plt.scatter(pred_lin, all_y, c='blue', alpha=0.5, marker='o', s=150, label =
u'без регуляризации')
    plt.scatter(pred_lasso, all_y, c='red', alpha=0.5, marker='s', s=150,
label='L1')
    plt.scatter(pred_ridge, all_y, c='green', alpha=0.5, marker='v', s=150,
label='L2')
    plt.ylabel(u'реальные значения')
    plt.xlabel(u'предсказанные значения')
    plt.legend(loc='upper left')
    plt.draw()
    plt.savefig('lin_regr_' + name + '.png', dpi=300)
    plt.show()
# Выбор данных и их нормализация
df_calc = all_df[[
        u'mark', u'base_s', u'base_max_int',
        u'Gauss.int_Ti', u'Gauss.int_Mo_Coh',
        u'Gauss.int_P', u'Gauss.int_Si', u'Gauss.int_Ta', u'Gauss.int_Zn',
        u'Gauss.int_Fe', u'Gauss.int_S', u'Gauss.int_K', u'Gauss.int_Mo',
        u'Gauss.int_Sr', u'Gauss.int_Ca', u'Gauss.int_Mn', u'Gauss.int_Cl',
        u'aver_gray', u'counters_num', u'counters_size'
    ]]
num_data = df_calc.ix[:, df_calc.columns != 'mark']
# normalisation without std
data_norm_r = (num_data - num_data.mean()) / (num_data.max()-num_data.min())
# normalisation with std (Z-conversation)
data_norm_z = (num_data - num_data.mean()) / (num_data.std())
labels = df_calc['mark']
# Кластеризация
names = ['data_norm_r', 'data_norm_z']
datas = [data_norm_r, data_norm_z]
```

```
for ind in xrange(len(datas)):
    X_raw = datas[ind]
    feature_names = labels
    transformer = random_projection.SparseRandomProjection(n_components = 2,
random_state = 12)
    # transformer = PCA(n_components=2)
    # transformer = manifold.MDS(n_components = 2, n_init = 1, max_iter = 100)
    # transformer = manifold.TSNE(n_components = 2, init = 'pca', random_state =
0)
    X_2d = transformer.fit_transform(X_raw)
    estimator = KMeans(n_clusters=len(Counter(labels)), init='k-means++',
n_init=10)
    estimator.fit(X_2d_norm)
    labels_t = estimator.labels_
    ftype_name = Counter(labels).keys()
    shifr = dict(zip(ftype_name, xrange(len(labels))))
    labels_class = np.array([shifr[x] for x in labels], dtype=int)
    # classifire with K-means
    classifier = KNeighborsClassifier()
    x, x_test, y, y_test = cross_validation.train_test_split(X_2d_norm,
labels_class, test_size=0.2, train_size=0.8)
    classifier.fit(x, y)
    for score in ['accuracy', 'precision_weighted', 'f1_weighted']:
            ridge_scoring = cross_validation.cross_val_score(classifier, x_test,
y_test, scoring = score, cv = 10)
            print score + ' mean: {}, std: {}, rel std:
{}'.format(ridge_scoring.mean(), ridge_scoring.std(), ridge_scoring.std())
    label_color = [colors[l] for l in labels_t]
    # plot clusters
    fig, ax = plt.subplots()
    plt.title(u"Кластеризация по марке\nслучайные проекции")
    ax.scatter(X_2d_norm[:,0], X_2d_norm[:,1], c=label_color, s=50)
    ax.scatter(estimator.cluster_centers_[:,0], estimator.cluster_centers_[:,1],
marker='*', s=150, c='g')
    shifr = dict(zip(ftype_name, xrange(len(labels))))
    inv_shifr = {v: k for k, v in shifr.iteritems()}
    shifr2 = {
        '12.52': u'МАФ\n12-52',
        '12.40.10': 'NP(S)\n12-40(10)',
        '0.20.20.5': 'NPK(S)\n0-20-20(5)',
        '4.30.15.16': 'NPK(S)\n4-30-15(16)',
        '16.16.8': 'NPK\n16-16-8'
    }
    title_font = {'fontname':'Arial', 'size':'16', 'color':'black',
'weight':'normal',
              'verticalalignment':'bottom'}
    _ = Counter(labels_t).keys()
    for ind2 in xrange(len(_)):
        k = _[ind2]
        x = X_2d_norm[:,1]
        y = X_2d_norm[:,0]
        plt.text(
            y[labels_t==k].mean(),
            x[labels_t==k].mean(),
            shifr2[str(inv_shifr[k])],
            horizontalalignment='center',
            bbox=dict(alpha=.5, edgecolor='w', facecolor='w'),
            **title_font
        )
    plt.draw()
    plt.savefig(names[ind] + '.png', dpi=300)
    plt.show()
```