

Круглый стол "Машинное обучение: эффективный инструмент
для познания физической картины мира"

18 ноября 2021 г., г. Москва

ОБРАБОТКА ДАННЫХ КОСВЕННЫХ ИЗМЕРЕНИЙ В СПЕКТРОСКОПИИ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

И.В. Исаев^{1,2}, О.Э. Сарманова³, С.А. Буриков³, Т.А. Доленко³,
К.А. Лаптинский¹, С.А. Доленко¹

¹ НИИ ядерной физики им. Д.В. Скобельцына, МГУ им. М.В. Ломоносова, г. Москва.

² Институт радиотехники и электроники им. В.А. Котельникова РАН, г. Москва.

² Физический факультет, МГУ им. М.В. Ломоносова, г. Москва.

*Исследование выполнено за счёт гранта Российского научного фонда
(проект № 19-11-00333)*

Введение

Прямые и косвенные измерения

- ❑ **Прямые измерения** — измерения, при которых искомое значение физической величины определяется непосредственно путём сравнения с мерой этой величины.
- ❑ **Косвенные измерения** — измерения, при которых искомое значение физической величины находится на основании известной зависимости между этой величиной и величинами, подвергаемыми прямым измерениям.

Введение

Решаемые задачи

- ❑ Определение качественного и количественного состава жидких сред:
 - Определение содержания и концентраций **ионов** растворов солей
 - ✓ Океанология, экологический мониторинг и т. д.
 - ✓ Контроль промышленных и сточных вод
 - ✓ Контроль качества минеральной воды
 - Определение концентраций **этанола и вредных примесей** водно-этанольных растворов
 - ✓ Контроль качества алкогольных напитков
 - Мониторинг выведения из организма **лекарственных препаратов и их метаболитов** с мочой
 - И др.

Введение

Методы определения концентраций веществ

❑ Методы определения концентраций веществ в растворах

- Химический анализ
- Хроматография
- Электрохимические методы
- Спектроскопические методы
- И т. д.

❑ Требования к методам определения концентраций

- Бесконтактность при использовании
- Определение в экспресс-режиме или в режиме реального времени
- Приемлемая точность

Введение

Традиционные методы

❑ Традиционные методы

- Химический анализ
- Хроматография
- Атомно-абсорбционная спектроскопия и др.

❑ Особенности методов

- Известна зависимость между наблюдаемой и определяемой величиной
- Высокая точность определения концентраций ионов
- Контактные
- Требуют значительное количество времени
- Каждый тест индивидуален для каждого вещества
- Требуют специальное лабораторное оборудование и реактивы
- Требуют высокую квалификацию персонала

Введение

Методы оптической спектроскопии

□ Методы оптической спектроскопии

- Спектроскопия КР
- Спектроскопия поглощения

□ Особенности методов

- Быстрые
- Бесконтактные
- Формы спектров КР и спектров поглощения растворов чувствительны к изменениям ионного состава и концентраций ионов.
- Отсутствует адекватная математическая модель, описывающая изменения формы спектров КР и поглощения.

Для обработки данных предлагается использовать
методы машинного обучения.

Спектроскопия КР растворов солей

Постановка задачи

Рассматриваемая задача заключается в идентификации и **определении концентраций** 10 ионов, содержащихся в многокомпонентных растворах 10 солей **по их КР спектрам**.

Определяемые ионы:

- Анионы: Cl^- , F^- , HCO_3^- , NO_3^- , SO_4^{2-}
- Катионы: K^+ , Li^+ , Mg^{2+} , Na^+ , NH_4^+

Соли:

- MgSO_4 , $\text{Mg}(\text{NO}_3)_2$, LiCl , LiNO_3 , NH_4F ,
 $(\text{NH}_4)_2\text{SO}_4$, KF , KHCO_3 , NaHCO_3 , NaCl .

Исследуемые растворы содержали от 2 до 10 ионов, содержащие от 1 до 5 солей в диапазоне концентраций 0-1.5 М, с шагом 0.15-0.25 М.

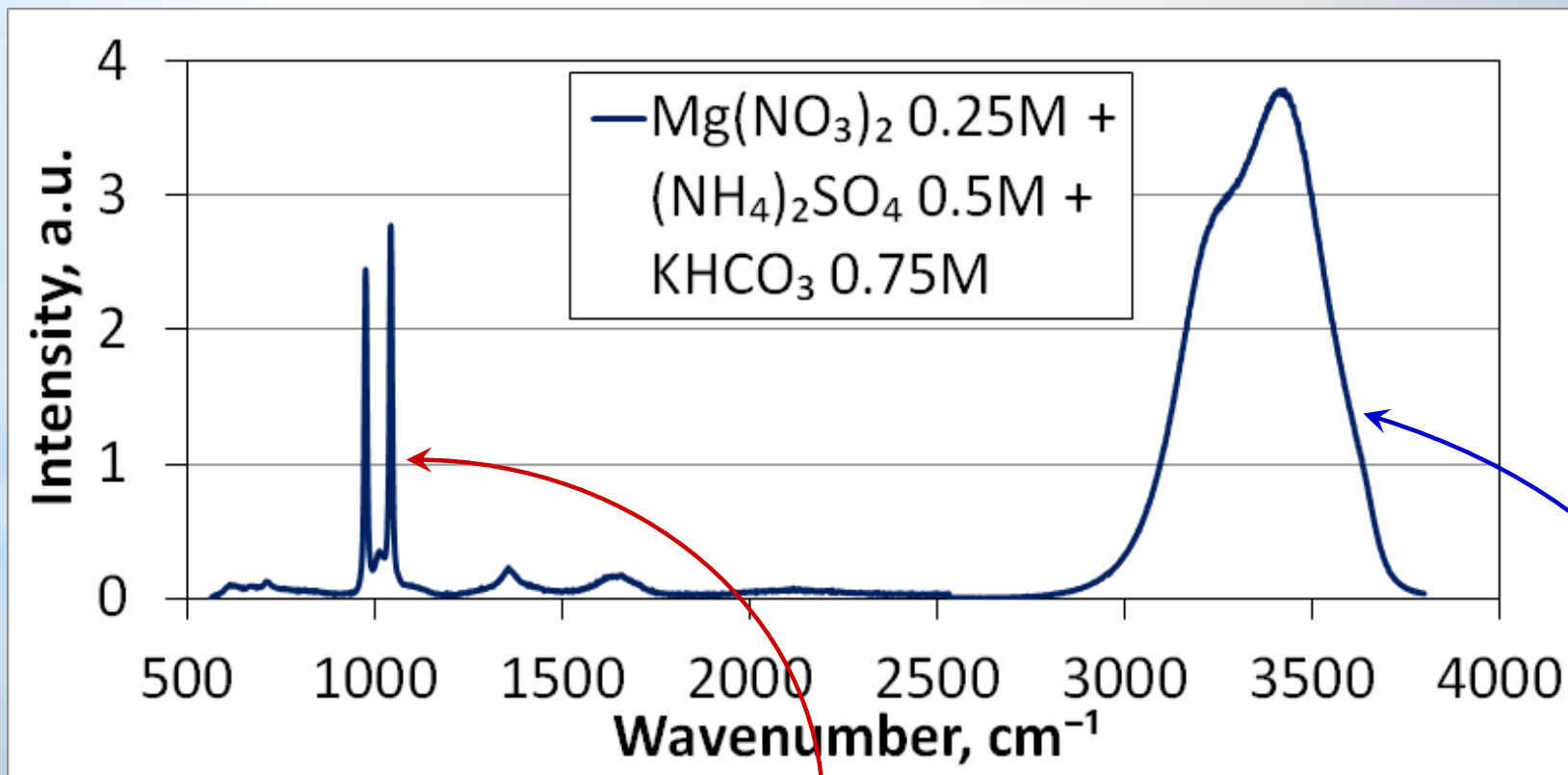
Спектроскопия КР растворов солей

Данные

- ❑ Набор данных получен экспериментально.
- ❑ Количество примеров: 4 445 примеров
- ❑ Размерность данных:
 - По входу 1 824 признаков
 - По выходу 10 параметров

Спектроскопия КР растворов солей

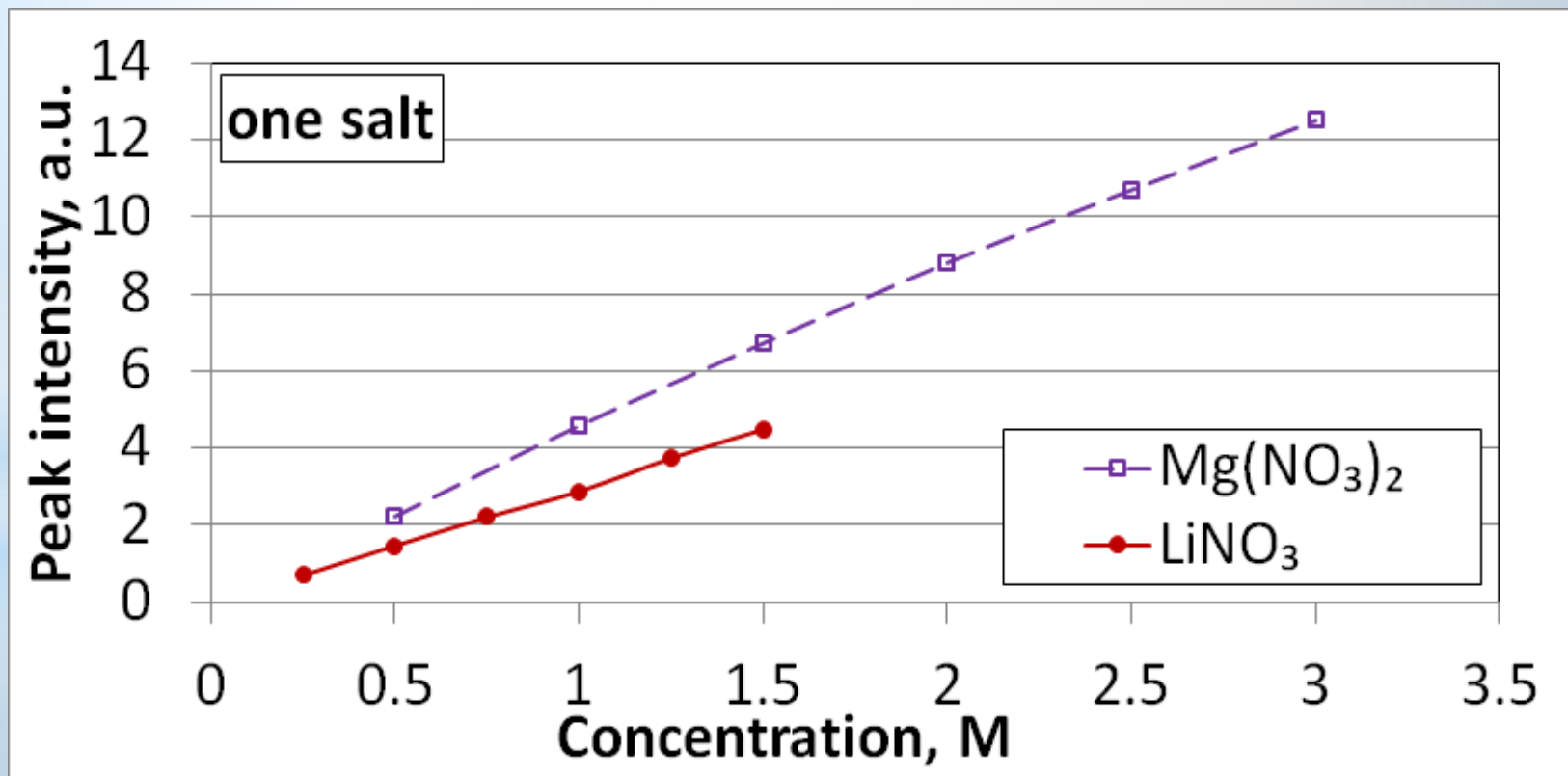
Пример спектра



Наблюдаются **характеристические линии** в левой части спектра, соответствующие **сложным ионам**, и **валентная полоса воды** в правой.

Спектроскопия КР растворов солей

Характеристические линии сложных ионов

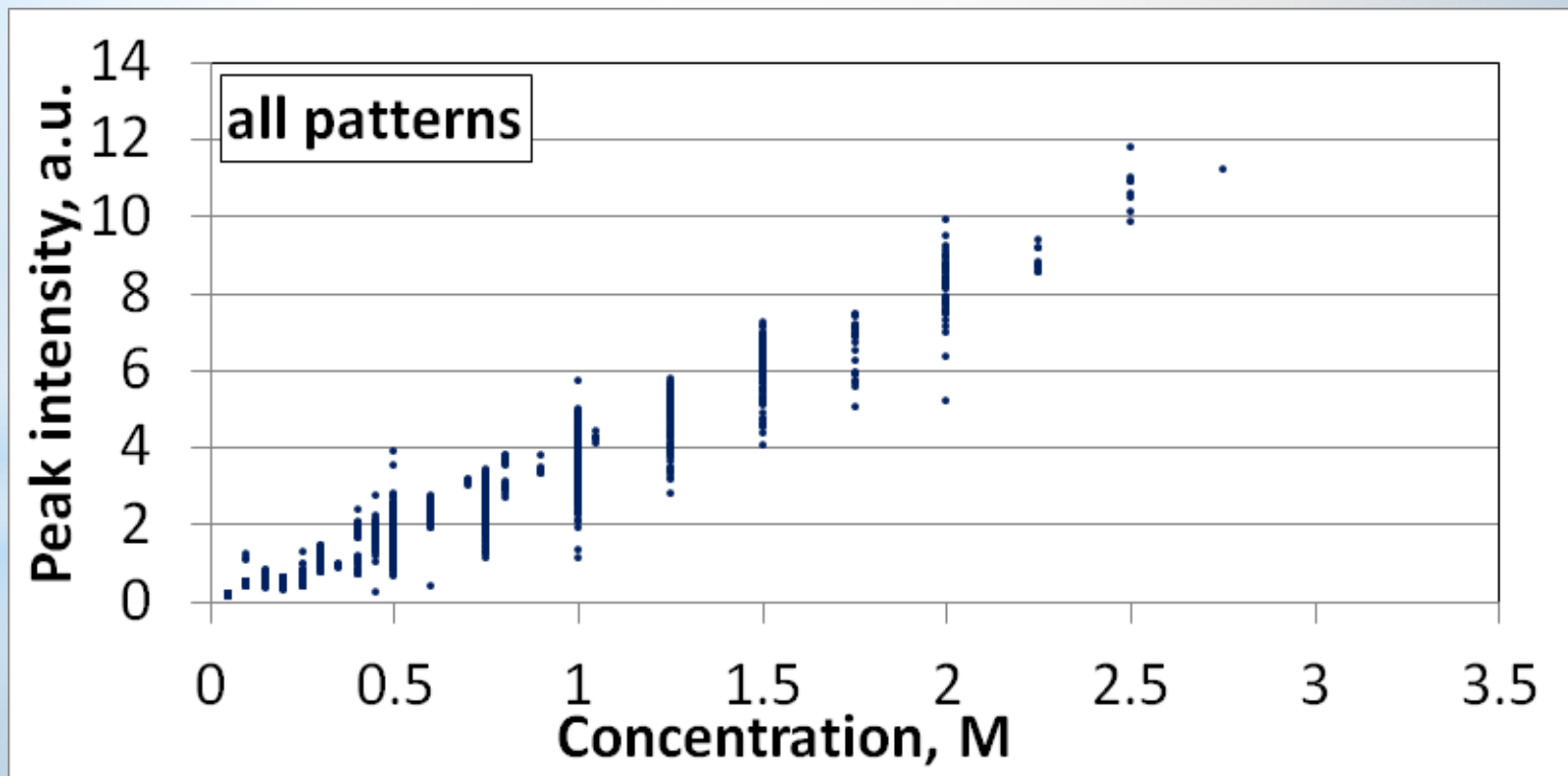


Интенсивность пиков, соответствующих сложным ионам, зависит от их концентрации в растворе.

Для раствора одной соли зависимость носит линейный характер.

Спектроскопия КР растворов солей

Характеристические линии сложных ионов

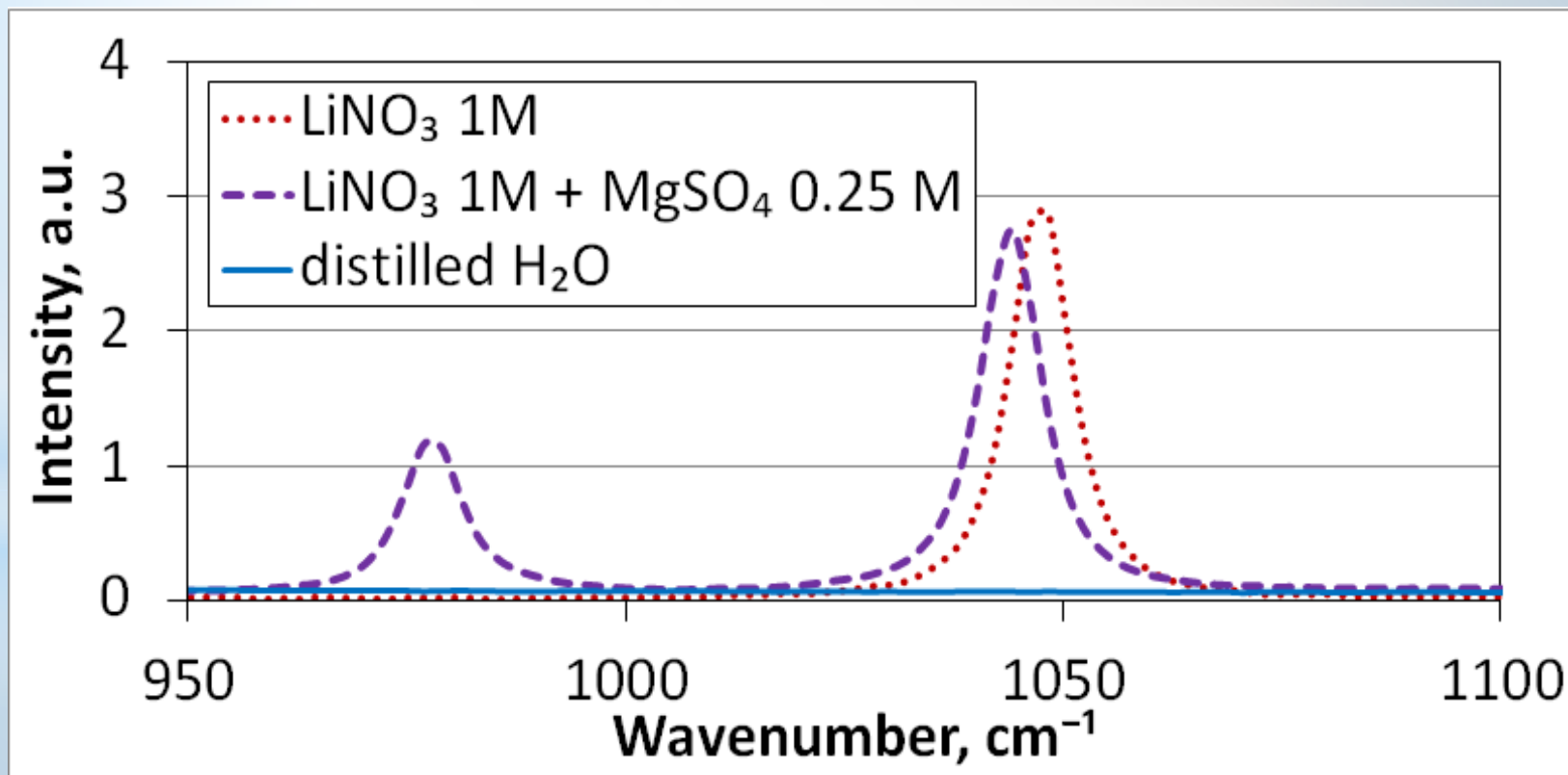


Интенсивность пиков, соответствующих сложным ионам, **зависит от их концентрации** в растворе.

При наличии **нескольких солей** в растворе **линейная зависимость нарушается**.

Спектроскопия КР растворов солей

Характеристические линии сложных ионов

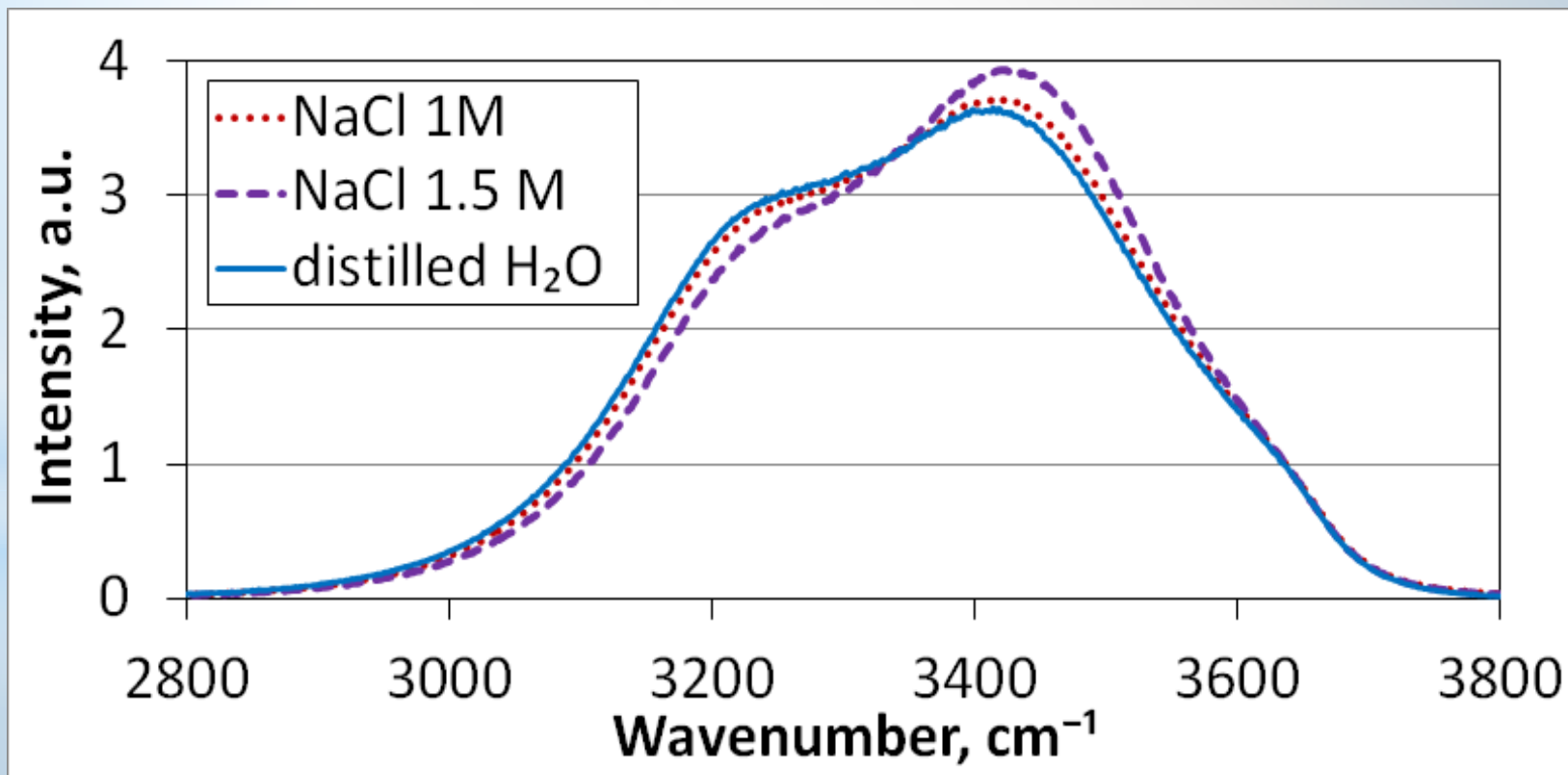


Интенсивность пиков, соответствующих сложным ионам, зависит от их концентрации в растворе.

При наличии нескольких солей в растворе линейная зависимость нарушается.

Спектроскопия КР растворов солей

Валентная полоса воды



Простые ионы не имеют собственных характеристических линий, но влияют на форму валентной полосы воды.

Спектроскопия водно-этанольных растворов

Постановка задачи

Рассматривается задача

бесконтактного контроля качества алкогольных напитков:

определение концентраций этанола и вредных примесей по их спектрам комбинационного рассеяния (КР).

Был получен массив спектров КР водных растворов этанола с вредными примесями, моделирующих водки различной крепости.

Определяемые компоненты:

- Этанол – 35..57 %
- Метанол – 0..24 %
- Этилацетат – 0..18 %
- Сивушное масло (моделируется смесью изоамилового и изопропилового спиртов) – 0..11.2 %

Спектроскопия водно-этанольных растворов

Данные

❑ Набор данных получен экспериментально.

❑ Количество примеров:

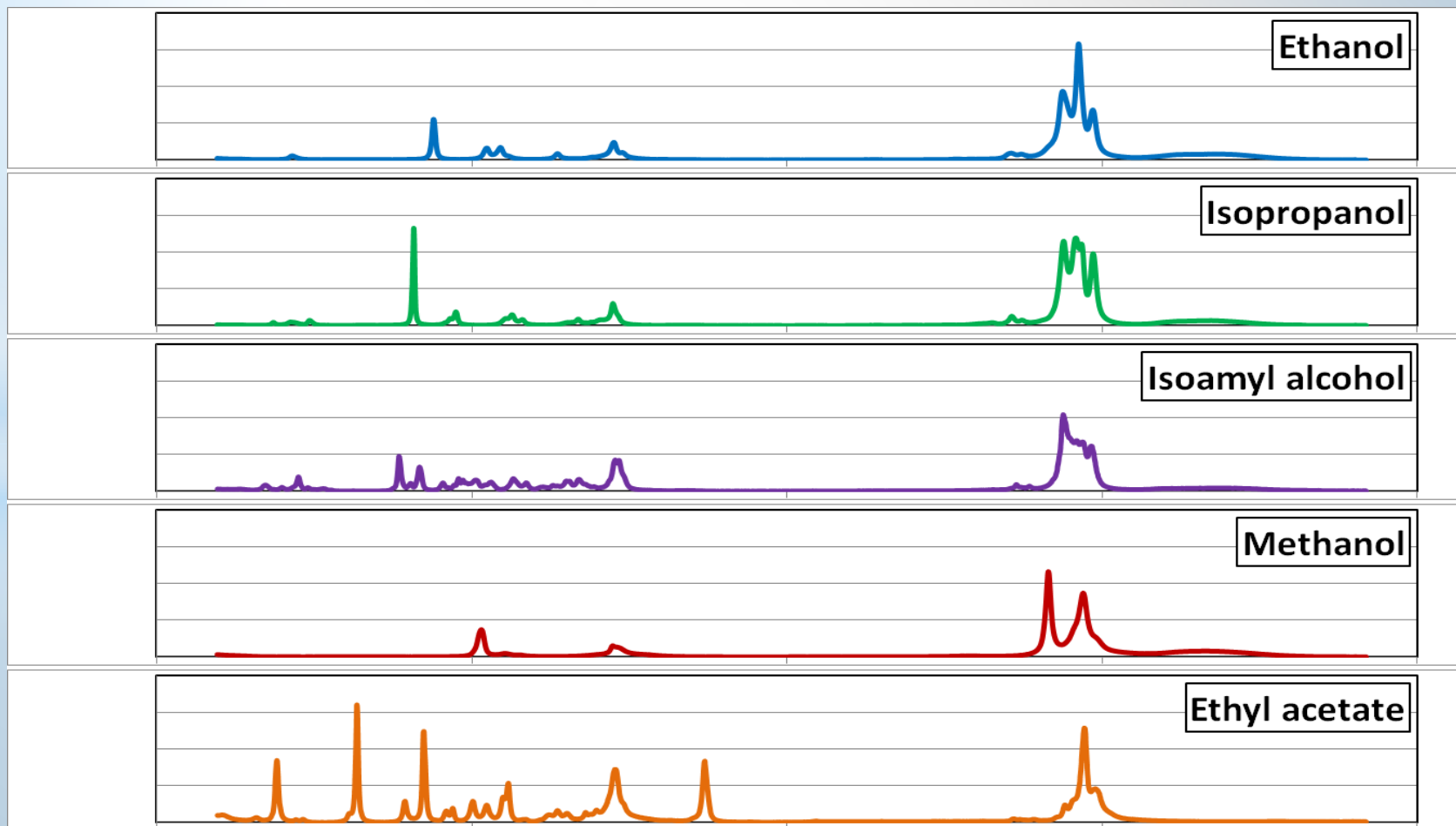
- Модельные растворы 4 046 примеров
- Реальные алкогольные напитки 69 примеров
 - ✓ водка
 - ✓ джин
 - ✓ текила
 - ✓ ликеры
 - ✓ самогон
 - ✓ и т. д.

❑ Размерность данных:

- По входу 2 048 признаков
- По выходу 4 параметра

Спектроскопия водно-этанольных растворов

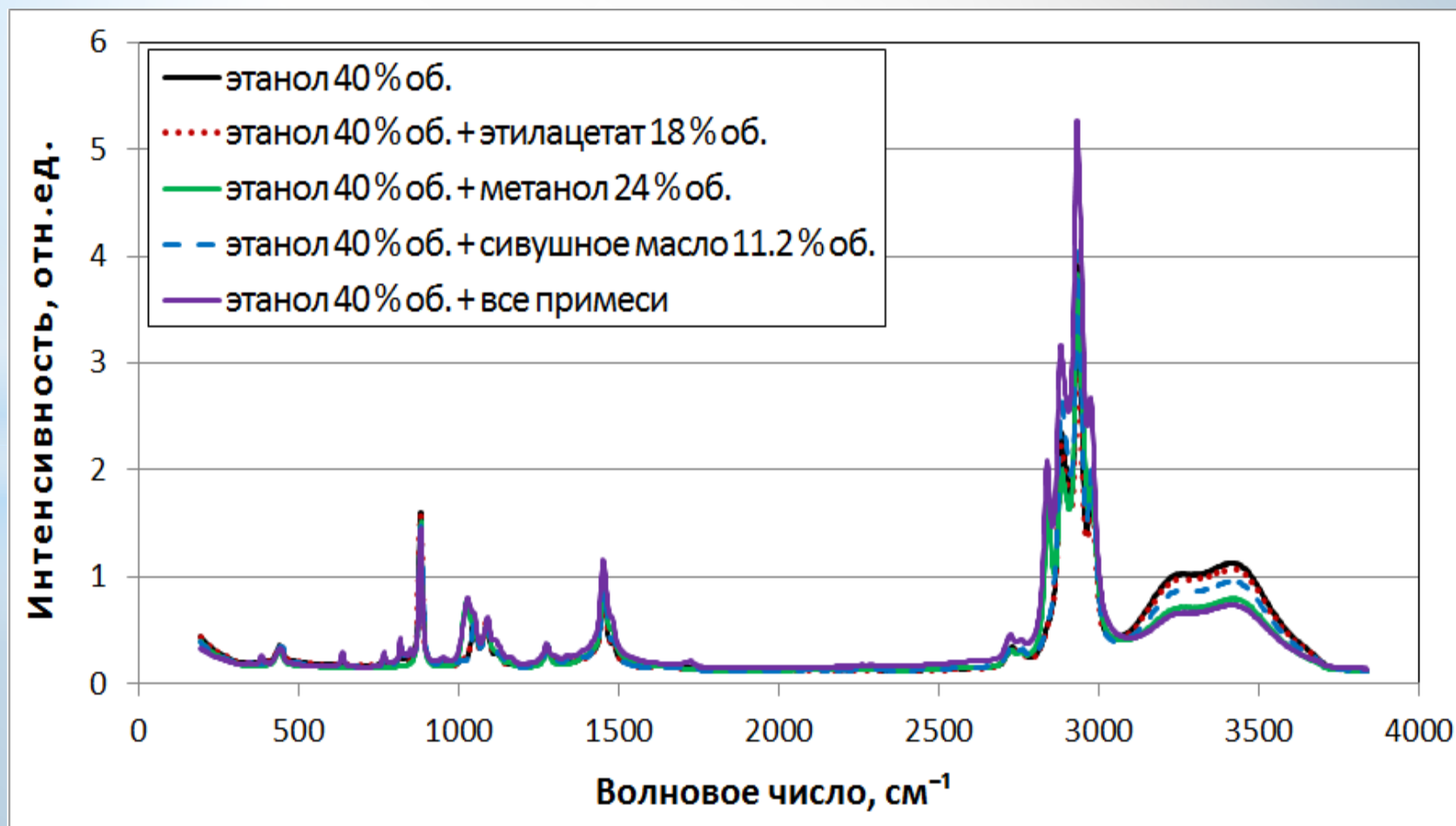
Спектры КР чистых компонентов



Характеристические линии компонентов
в значительной мере перекрываются

Спектроскопия водно-этанольных растворов

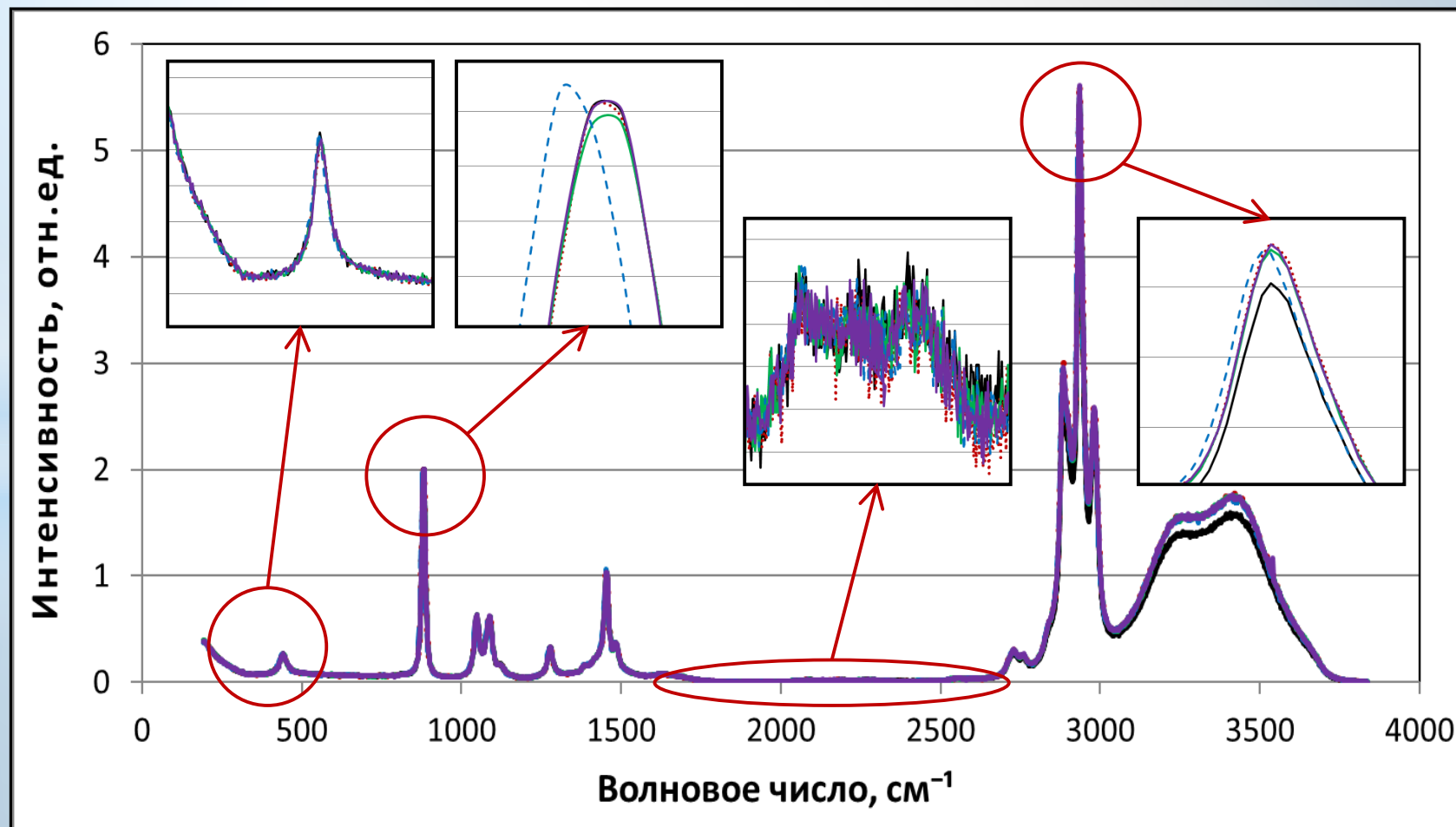
Спектры КР водно-этанольных растворов



Спектры КР водно-этанольных растворов оказываются
близки друг к другу

Виды искажений данных спектроскопии

Примеры спектров, содержащие искажения



Спектры КР одного образца могут **отличаться** друг от друга

Вид искажений данных спектроскопии

Описание искажений в данных

Типы искажений:

Отклонения
концентраций компонентов

Случайный шум
в определении интенсивности
каналов спектра

Сдвиг каналов спектра

Пьедестал

Растяжение или сжатие спектра

Источник искажений:

Неточности при приготовлении
растворов

Темновой шум
и погрешность определения
сигнала ПЗС-детектором

Изменение
юстировки установки

Рассеяние света
на неоднородностях среды

Изменение мощности лазера
или коэффициента
поглощения контейнера

Особенности задач оптической спектроскопии

Особенности задач

Особенности **задач** оптической спектроскопии:

- ☐ Не имеют ни аналитического, ни прямого численного решения
- ☐ Нелинейные
- ☐ Имеет высокую размерность
 - по входу
 - по выходу (многопараметрическая)
- ☐ Имеется мультиколлинеарность входных признаков
- ☐ Являются обратными задачами
 - Некорректные и плохо обусловленные
 - Обладают высокой чувствительностью решения к шумам в данных

Особенности задач оптической спектроскопии

Особенности данных

Особенности **данных** оптической спектроскопии:

- ❑ Содержат шум различной природы
- ❑ Имеют малую репрезентативность:
 - Малая размерность тренировочной выборки
 - Малая репрезентативность самих данных:
 - ✓ Наборы данных не содержат все возможные сочетания компонентов
 - ✓ Наборы данных не содержат все возможные сочетания концентраций компонентов.

Использование методов машинного обучения

Особенности применения

Особенности применения методов машинного обучения для решения обратных задач спектроскопии:

- ❑ Предобработка данных
- ❑ Использование простых нелинейных методов:
 - Персептрон с небольшим количеством скрытых слоев
- ❑ Снижение размерности:
 - По входу: отбор существенных признаков
 - По выходу: автономное и групповое определение параметров
- ❑ Использование специальных подходов для повышения устойчивости решения к шумам в данных
 - Групповое определение параметров (“Multi-task learning”)
 - Добавление шума при обучении

Использование методов машинного обучения

Способы предобработки данных

Типы искажений:

Отклонения
концентраций компонентов



Случайный шум
в определении интенсивности
каналов спектра



Сдвиг каналов спектра



Пьедестал



Растяжение или сжатие спектра



Способы предобработки:

Альтернативные методы
контроля

Измерение
нескольких спектров
с последующим усреднением

Контроль по линиям
известных компонентов

Вычитание пьедестала

Нормировка
на валентную полосу воды
Для спиртов **не применима**

Использование методов машинного обучения

Снижение выходной размерности задачи

❑ Автономное определение:

- Решение отдельных задач с одним выходом
- Построение отдельной сети для каждого из определяемых параметров

❑ Групповое определение:

- Объединение параметров в группы
- Одновременное определение параметров (и построение одной НС) внутри каждой группы
- Использовались следующие принципы группировки параметров:

✓ Простые ионы	6	параметров
✓ Сложные ионы	4	параметра
✓ Катионы	5	параметров
✓ Анионы	5	параметров

Использование методов машинного обучения

Метод добавления шума при обучении НС

- ❑ **Оптимальным режимом** обучения с добавлением шума оказался такой, когда обучение проводилось на тренировочном наборе, который содержал шум, а остановка обучения осуществлялась по валидационному набору без добавления шума.
 - В этом случае **качество** решения было **выше**, а **время** обучения - **ниже**.
- ❑ Добавление шума производилось перед обучением.
- ❑ Каждый пример исходных наборов был представлен в нескольких зашумленных реализациях.
- ❑ Сети, обученные на тренировочном наборе с фиксированным уровнем шума определенного типа, применялись к тестовым наборам с различными уровнями шума того же типа

Задача “10 ионов”. Случайный шум.

Процедура добавления шума

- ❑ Использовался **аддитивный Гауссов (*agn*)** и **аддитивный равномерный (*aun*)** шум.
- ❑ Значение каждого входного признака x_i преобразовывалось следующим образом:
 - $x_i^{agn} = x_i + norm_inv(random, \mu = 0, \sigma = noise_level) \cdot \max(x_i)$
 - $x_i^{aun} = x_i + (1 - 2 \cdot random) \cdot noise_level \cdot \max(x_i)$

✓ $random$	–	случайное число в диапазоне 0..1
✓ $norm_inv$	–	функция, возвращающая обратное нормальное распределение
✓ $noise_level$	–	уровень шума
✓ $\max(x_i)$	–	максимальное значение интенсивности в i - м канале спектра

по всему набору данных
- ❑ Использовались уровни шума 1%, 3%, 5%, 10%, 20%.
- ❑ Каждый пример исходных наборов был представлен в 10 зашумленных реализациях.

Задача “10 ионов”. Случайный шум.

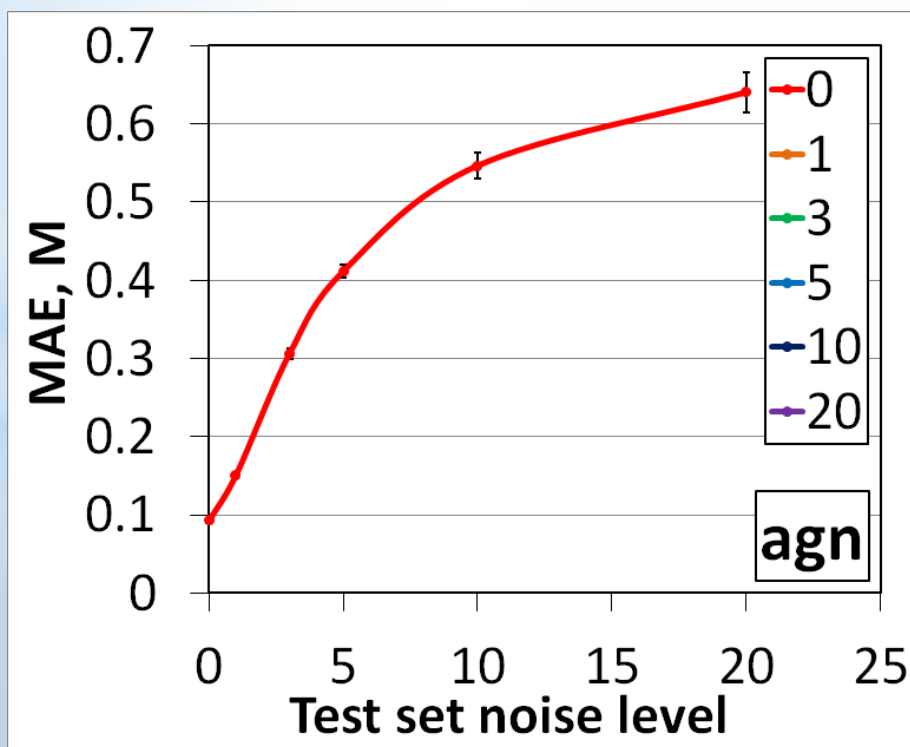
Параметры нейронных сетей

- ❑ Топология: **персептрон** с 3 скрытыми слоями.
- ❑ Количество нейронов в скрытых слоях: 64, 32, 16.
- ❑ Передаточная функция:
 - скрытые слои — сигмоидальная
 - выходной слой — линейная
- ❑ Остановка обучения:
 - 1000 эпох без улучшения результата на валидационном наборе.
- ❑ Каждая сеть тренировалась по 5 раз с различными случайными комплектами начальных весов с последующим усреднением статистик результатов применения.

Задача “10 ионов”. Случайный шум.

Результаты

Зависимость качества решения (MAE) для иона Cl^- от уровня шума (agn) в тестовом наборе.



Линии – автономное определение.

Маркеры – групповое определение.

Красный цвет – без шума

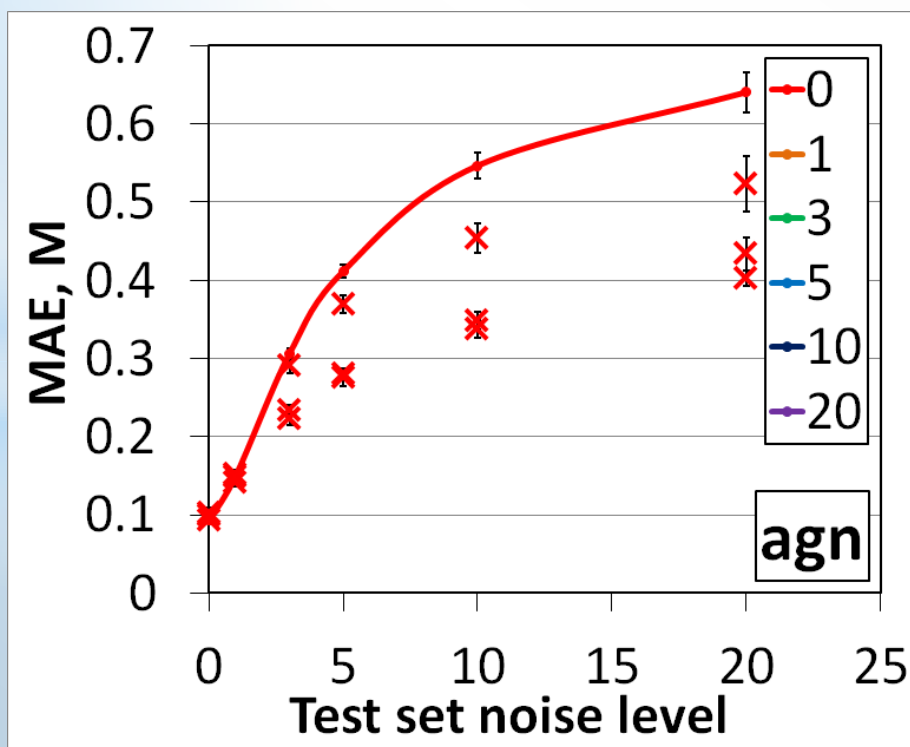
Другие цвета – метод обучения с добавлением шума

Качество нейросетевого решения ОЗ падает с увеличением уровня шума в тестовой выборке

Задача “10 ионов”. Случайный шум.

Результаты

Зависимость качества решения (MAE) для иона Cl^- от уровня шума (agn) в тестовом наборе.



Линии – автономное определение.

Маркеры – групповое определение.

Красный цвет – без шума

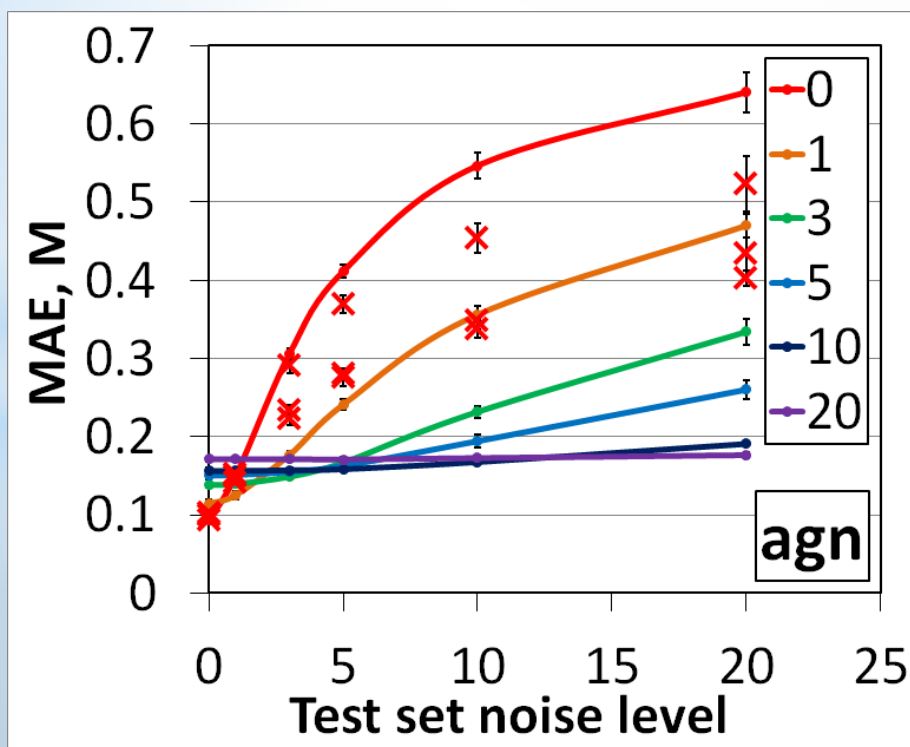
Другие цвета – метод обучения с добавлением шума

Групповое определение позволяет повысить устойчивость нейросетевого решения ОЗ к шумам в данных

Задача “10 ионов”. Случайный шум.

Результаты

Зависимость качества решения (MAE) для иона Cl^- от уровня шума (agn) в тестовом наборе.



Линии – автономное определение.

Маркеры – групповое определение.

Красный цвет – без шума

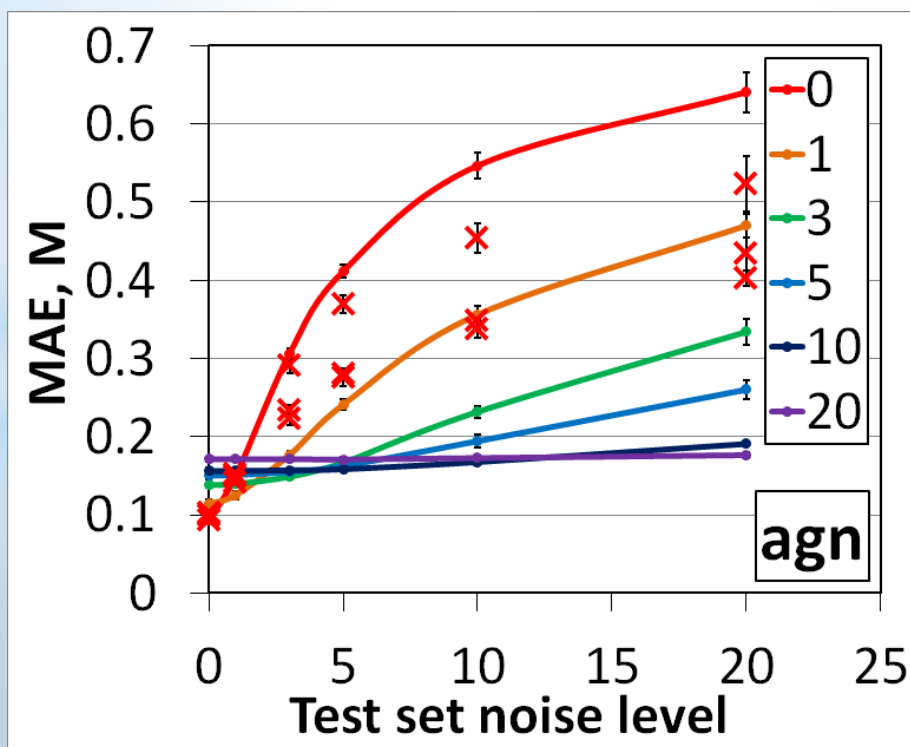
Другие цвета – метод обучения с добавлением шума

Чем выше уровень шума в тренировочном наборе, тем хуже сеть работает с данными без шума.

Задача “10 ионов”. Случайный шум.

Результаты

Зависимость качества решения (MAE) для иона Cl^- от уровня шума (agn) в тестовом наборе.



Линии – автономное определение.

Маркеры – групповое определение.

Красный цвет – без шума

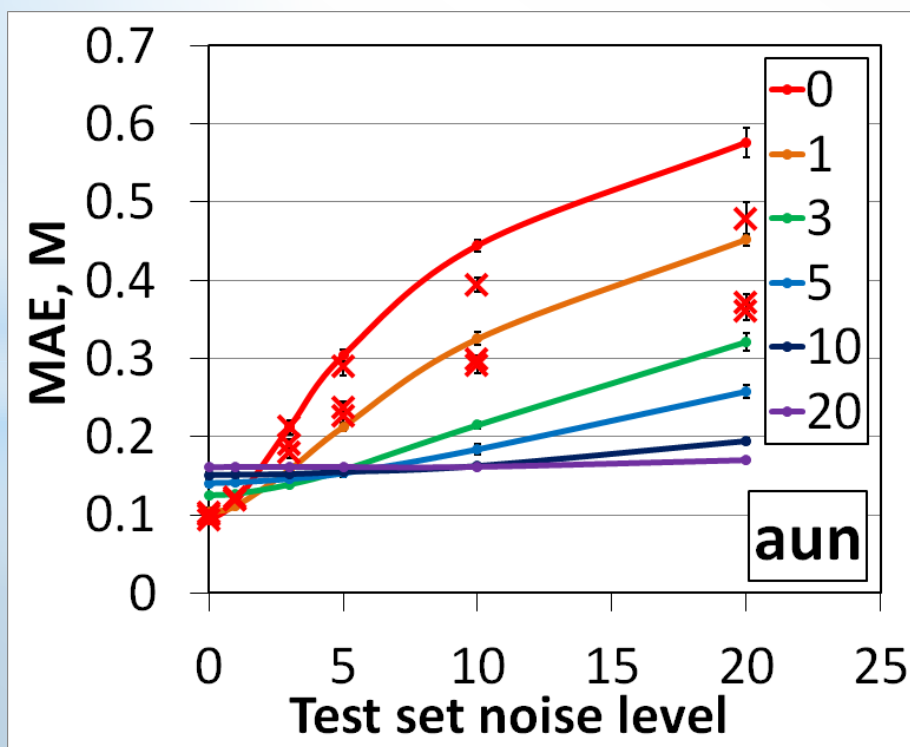
Другие цвета – метод обучения с добавлением шума

Чем выше уровень шума в тренировочном наборе, тем медленнее деградирует качество решения с увеличением уровня шума в тестовом наборе.

Задача “10 ионов”. Случайный шум.

Результаты

Зависимость качества решения (MAE) для иона Cl^- от уровня шума (aun) в тестовом наборе.



Линии – автономное определение.

Маркеры – групповое определение.

Красный цвет – без шума

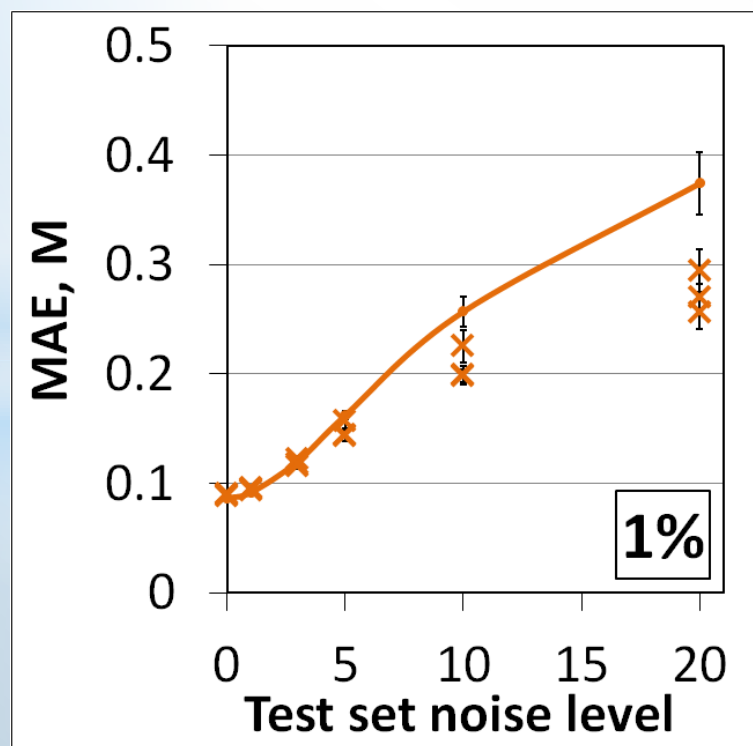
Другие цвета – метод обучения с добавлением шума

Данный эффект наблюдается для всех определяемых ионов и для всех типов и статистик шума.

Задача “10 ионов”. Случайный шум.

Результаты

Зависимость качества решения (MAE) для иона K^+ от уровня шума (agn) в тестовом наборе.



Линии – автономное определение.

Маркеры – групповое определение.

Красный цвет – без шума

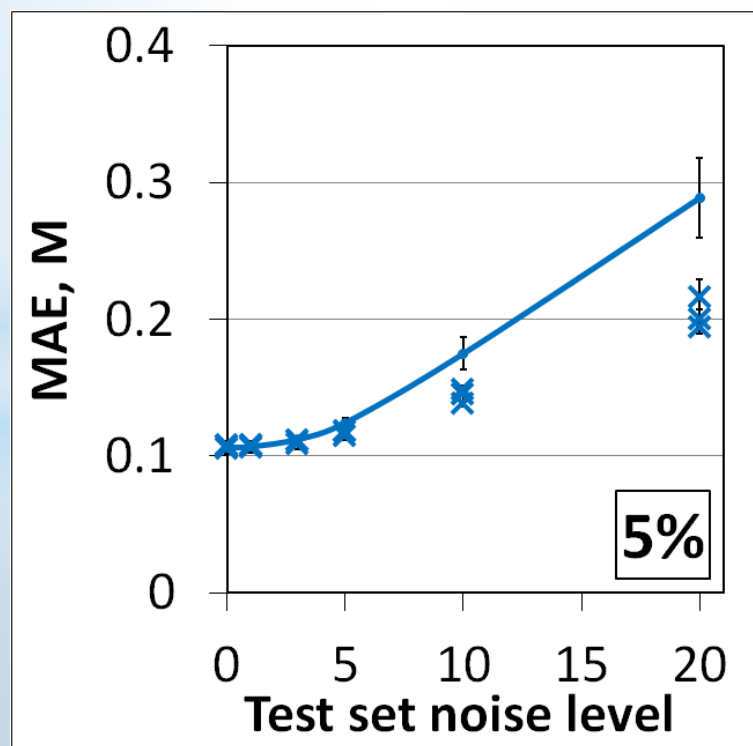
Другие цвета – метод обучения с добавлением шума

Совместное использование методов позволяет повысить устойчивость решения по сравнению с автономным определением с добавлением шума.

Задача “10 ионов”. Случайный шум.

Результаты

Зависимость качества решения (MAE) для иона K^+ от уровня шума (agn) в тестовом наборе.



Линии – автономное определение.

Маркеры – групповое определение.

Красный цвет – без шума

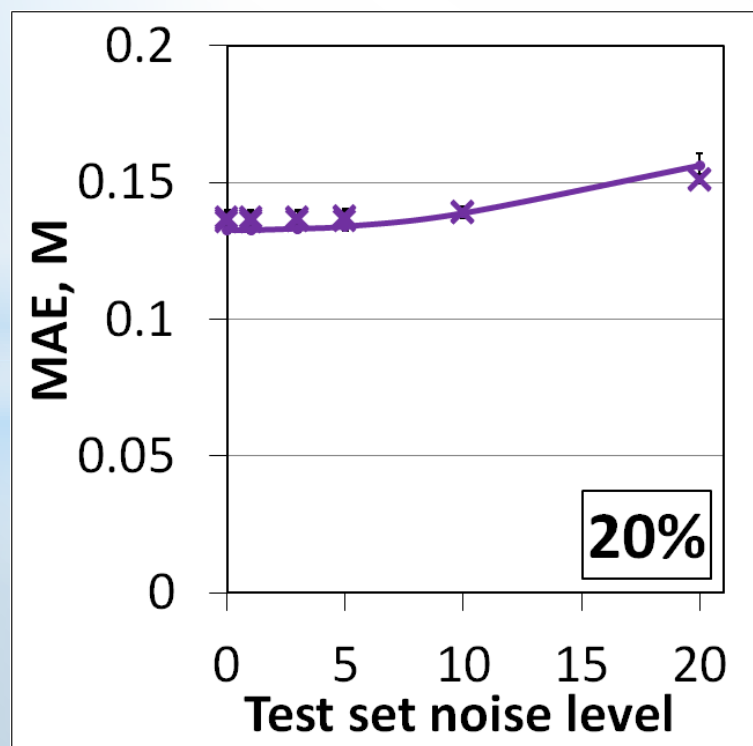
Другие цвета – метод обучения с добавлением шума

Эффект от совместного использования более заметен при высоких уровнях шума в тестовом наборе и низких уровнях шума в тренировочном наборе.

Задача “10 ионов”. Случайный шум.

Результаты

Зависимость качества решения (MAE) для иона K^+ от уровня шума (agn) в тестовом наборе.



Линии – автономное определение.

Маркеры – групповое определение.

Красный цвет – без шума

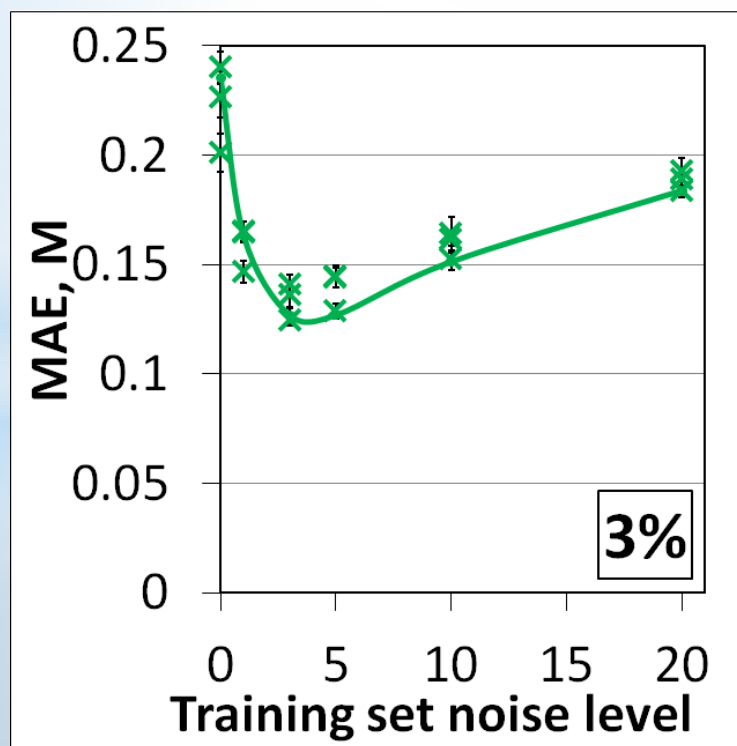
Другие цвета – метод обучения с добавлением шума

Эффект от совместного использования более заметен при высоких уровнях шума в тестовом наборе и низких уровнях шума в тренировочном наборе.

Задача “10 ионов”. Случайный шум.

Результаты

Зависимость качества решения (MAE) для иона NH_4^+ от уровня шума (aun) в тренировочном наборе.



Линии – автономное определение.

Маркеры – групповое определение.

Красный цвет – без шума

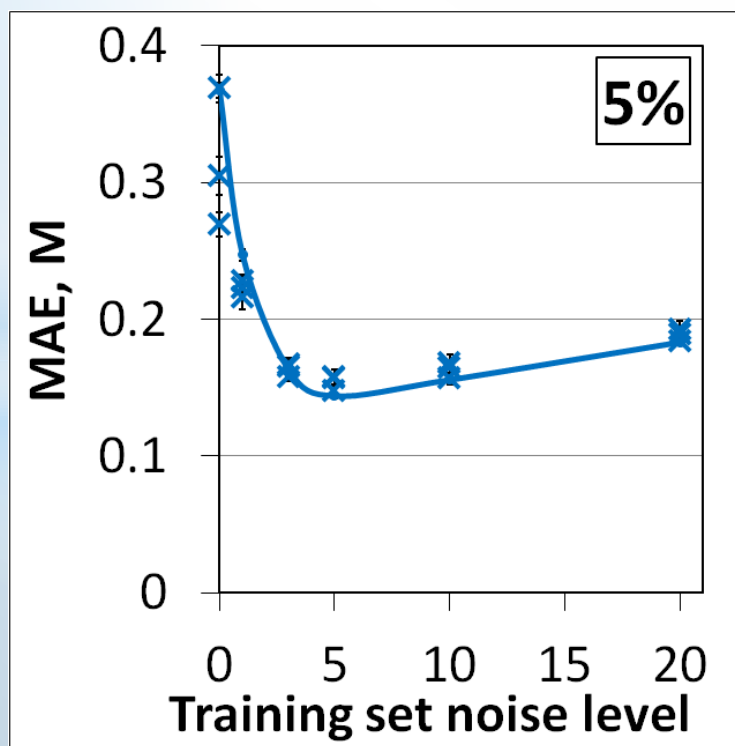
Другие цвета – метод обучения с добавлением шума

Наилучшее качество решения для метода обучения с шумом наблюдается когда уровень шума в тренировочном наборе и тестовом наборе совпадают.

Задача “10 ионов”. Случайный шум.

Результаты

Зависимость качества решения (MAE) для иона NH_4^+ от уровня шума (aun) в тренировочном наборе.



Линии – автономное определение.

Маркеры – групповое определение.

Красный цвет – без шума

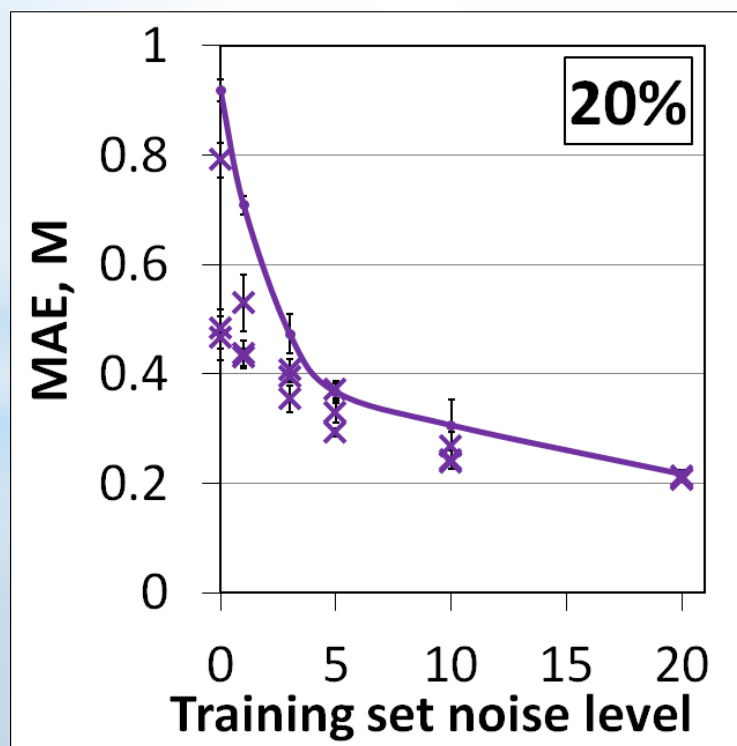
Другие цвета – метод обучения с добавлением шума

Наилучшее качество решения для метода обучения с шумом наблюдается когда уровень шума в тренировочном наборе и тестовом наборе совпадают.

Задача “10 ионов”. Случайный шум.

Результаты

Зависимость качества решения (MAE) для иона NH_4^+ от уровня шума (aun) в тренировочном наборе.



Линии – автономное определение.

Маркеры – групповое определение.

Красный цвет – без шума

Другие цвета – метод обучения с добавлением шума

Наилучшее качество решения для метода обучения с шумом наблюдается когда уровень шума в тренировочном наборе и тестовом наборе совпадают.

Задача “10 ионов”. Сдвиг каналов спектра.

Процедура добавления шума

- ❑ Каждый пример исходных наборов был представлен в 5 реализациях:
 - Без сдвига каналов спектра
 - Со сдвигом на 1 или 2 канала вправо
 - Со сдвигом на 1 или 2 канала влево

Задача “10 ионов”. Сдвиг каналов спектра.

Параметры нейронных сетей

- ❑ Топология: **персептрон** с 3 скрытыми слоями.
- ❑ Количество нейронов в скрытых слоях: 64, 32, 16.
- ❑ Передаточная функция:
 - скрытые слои — сигмоидальная
 - выходной слой — линейная
- ❑ Остановка обучения:
 - 1000 эпох без улучшения результата на валидационном наборе.
- ❑ Каждая сеть тренировалась по 5 раз с различными случайными комплектами начальных весов с последующим усреднением статистик результатов применения.

Задача “10 ионов”. Сдвиг каналов спектра.

Применение нейронных сетей

☐ No shift to no shift

- тренировочный набор не содержал искажения,
- тестовый набор не содержал искажения.

☐ No shift to shift

- тренировочный набор не содержал искажения,
- тестовый набор содержал искажения.

☐ Shift to no shift

- тренировочный набор содержал искажения,
- тестовый набор не содержал искажения.

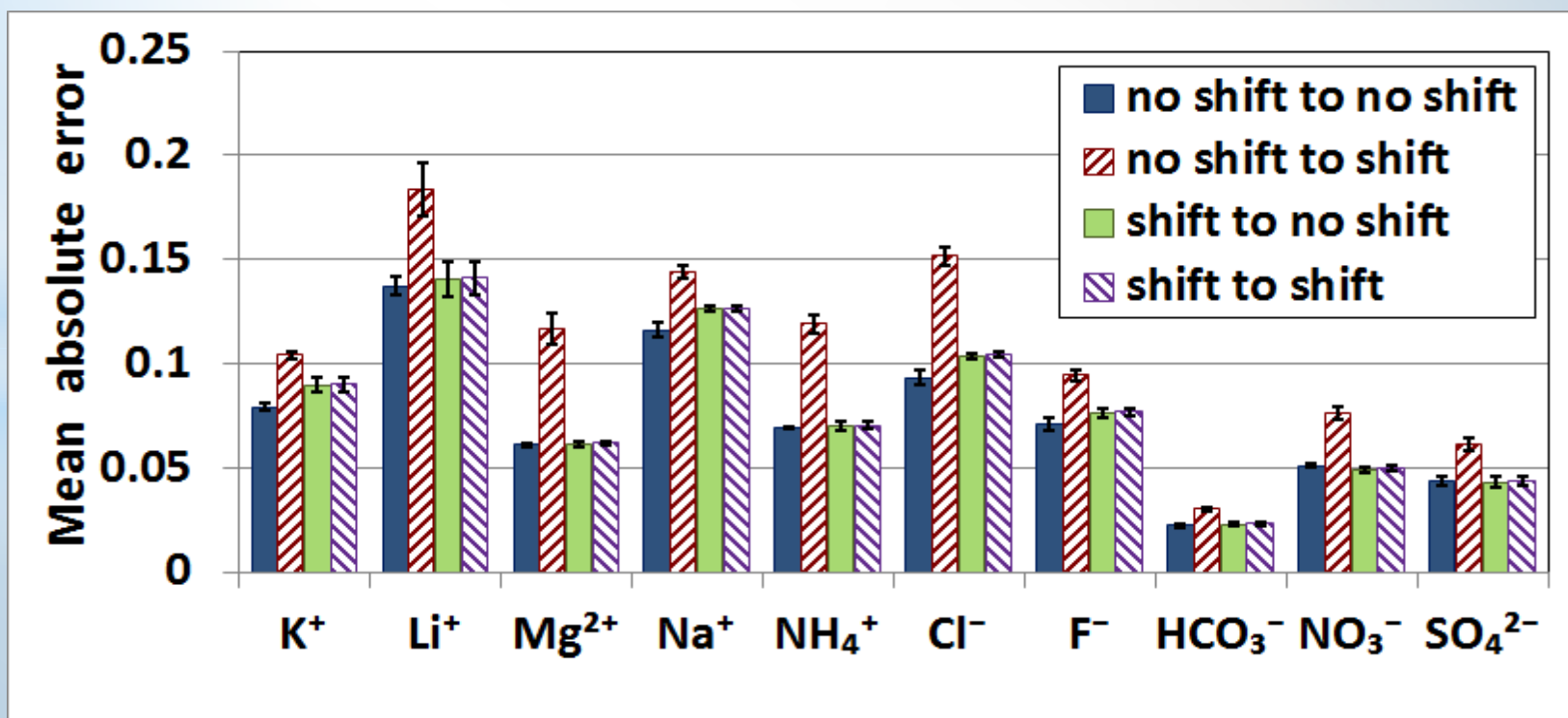
☐ Shift to shift

- тренировочный набор содержал искажения,
- тестовый набор содержал искажения.

Задача “10 ионов”. Сдвиг каналов спектра.

Результаты

Зависимость качества решения (MAE) для каждого иона от способа применения нейронных сетей.



Обучение с добавлением примеров, содержащих сдвиг каналов спектра, позволяет повысить устойчивость решения к данному типу искажений.

Задача “Спирты”. Растяжение/сжатие спектров.

Процедура добавления шума

- ❑ Использовался мультипликативный Гауссов (*mgn*) и мультипликативный равномерный (*mun*) шум.
- ❑ Значение вектора входных признаков X преобразовывалось следующим образом:
 - $X^{mgn} = X \cdot (1 + norm_inv(random, \mu = 0, \sigma = noise_level))$
 - $X^{mun} = X \cdot (1 + (1 - 2 \cdot random) \cdot noise_level)$
 - ✓ *random* – случайное число в диапазоне 0..1
 - ✓ *norm_inv* – функция, возвращающая обратное нормальное распределение
 - ✓ *noise_level* – уровень шума
- ❑ Использовались уровни шума 1%, 3%, 5%, 10%, 20%.
- ❑ Каждый пример исходных наборов был представлен в 10 зашумленных реализациях.

Задача “Спирты”. Растяжение/сжатие спектров.

Параметры нейронных сетей

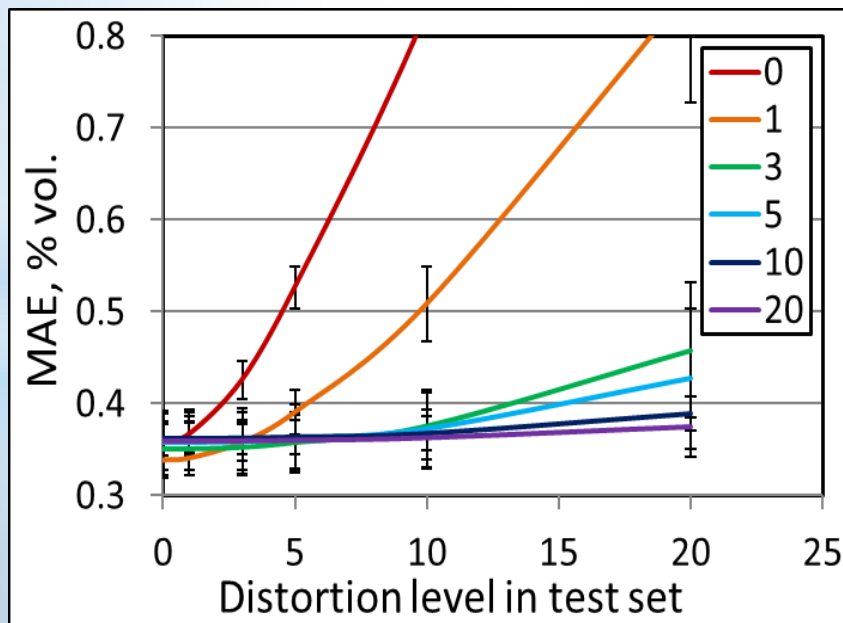
- ❑ Топология: **персептрон** с 1 скрытым слоем.
- ❑ Количество нейронов в скрытых слоях: 32.
- ❑ Передаточная функция:
 - скрытый слой — сигмоидальная
 - выходной слой — линейная
- ❑ Остановка обучения:
 - 500 эпох без улучшения результата на валидационном наборе.
- ❑ Каждая сеть тренировалась по 5 раз с различными случайными комплектами начальных весов с последующим усреднением статистик результатов применения.

Задача “Спирты”. Растяжение/сжатие спектров.

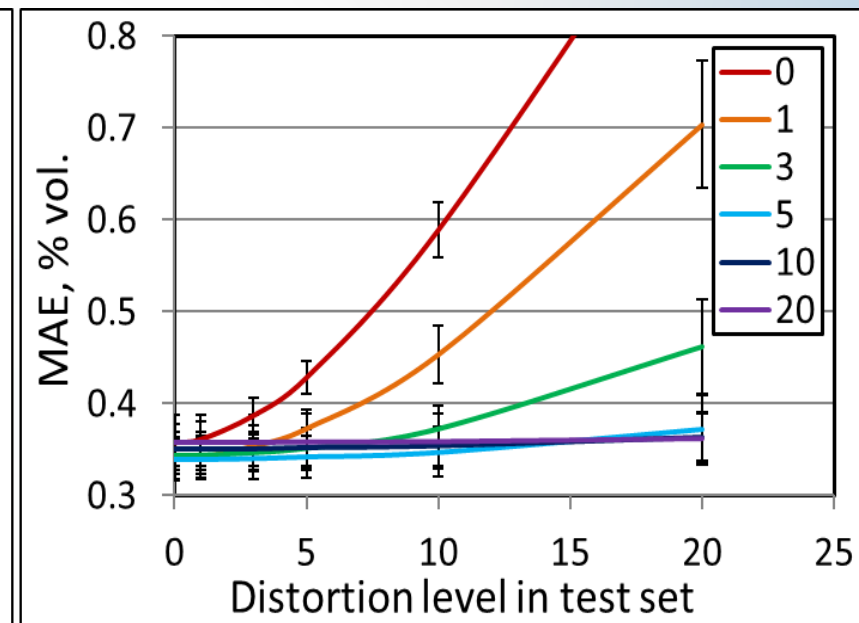
Набор данных модельных растворов

Этанол

мультипликативный Гауссов



мультипликативный равномерный

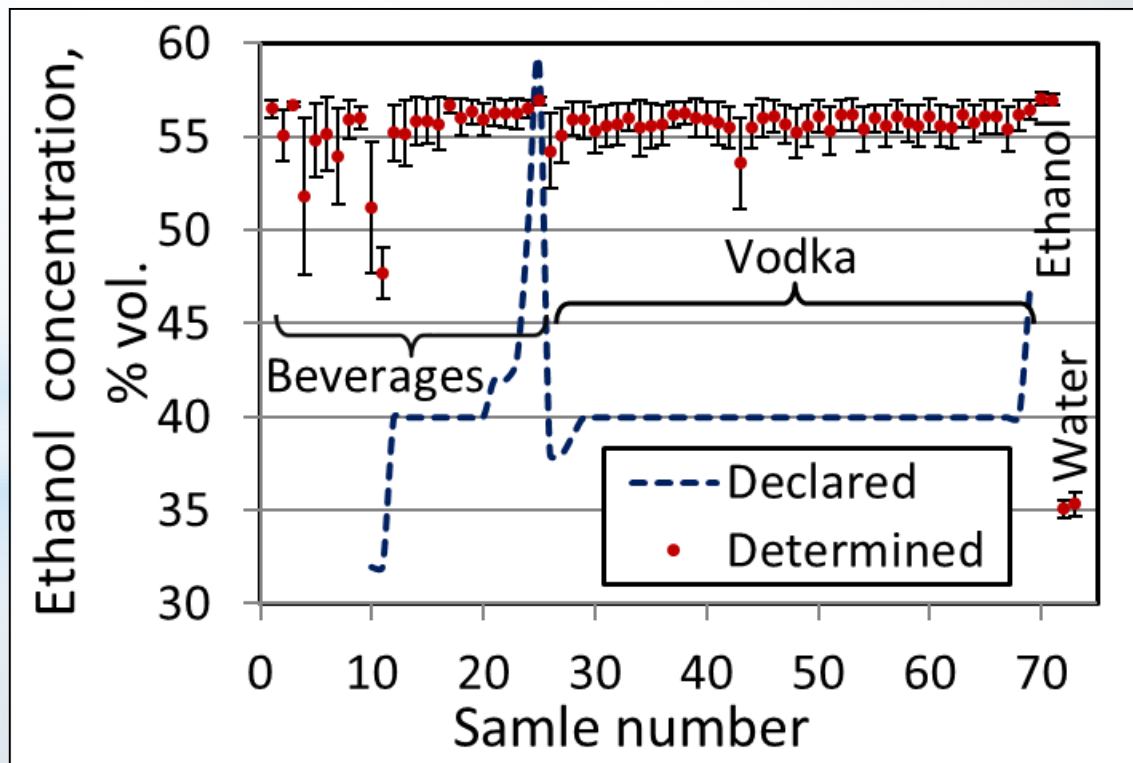


Чем выше уровень шума в тренировочном наборе, тем хуже сеть работает с данными без шума, но тем медленнее деградирует качество решения с увеличением уровня шума в тестовом наборе.

Задача “Спирты”. Растяжение/сжатие спектров.

Набор данных реальных алкогольных напитков

Сети, обученные **без добавления** шума

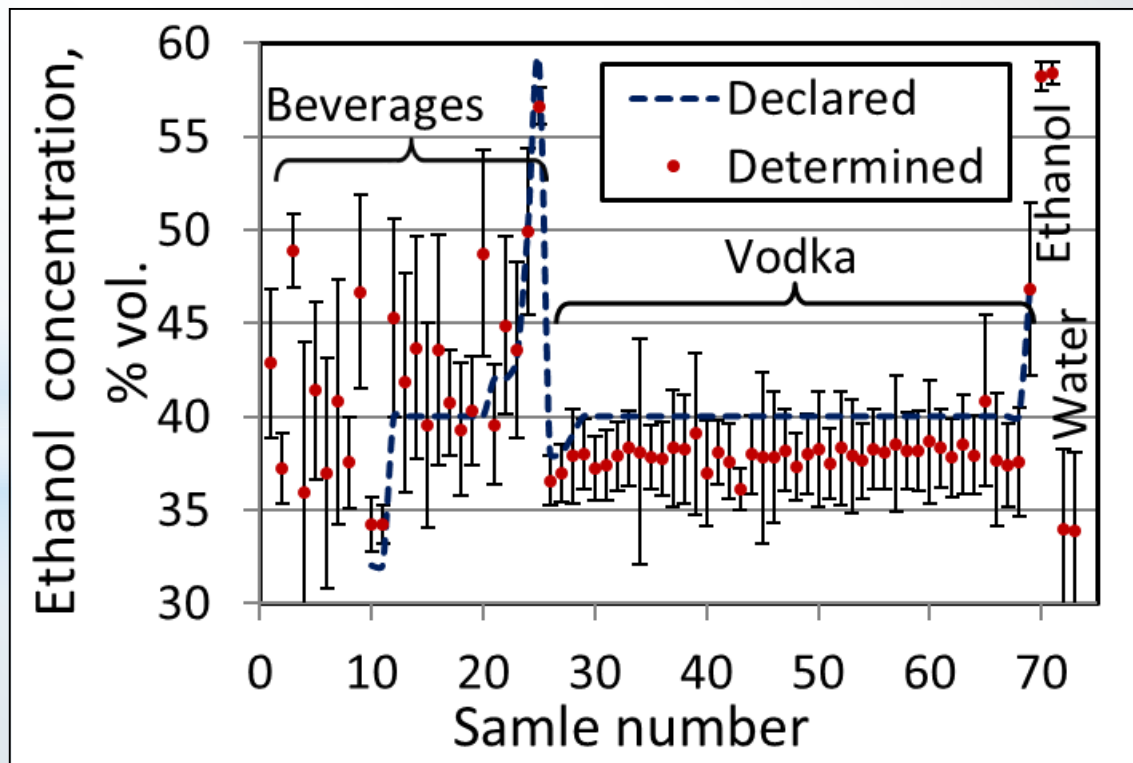


Результаты практически для всего набора данных **ушли в насыщение** - показали нижний или верхний уровень концентраций в тренировочном наборе

Задача “Спирты”. Растяжение/сжатие спектров.

Набор данных реальных алкогольных напитков

Сети, обученные с добавлением 20% Гауссова шума

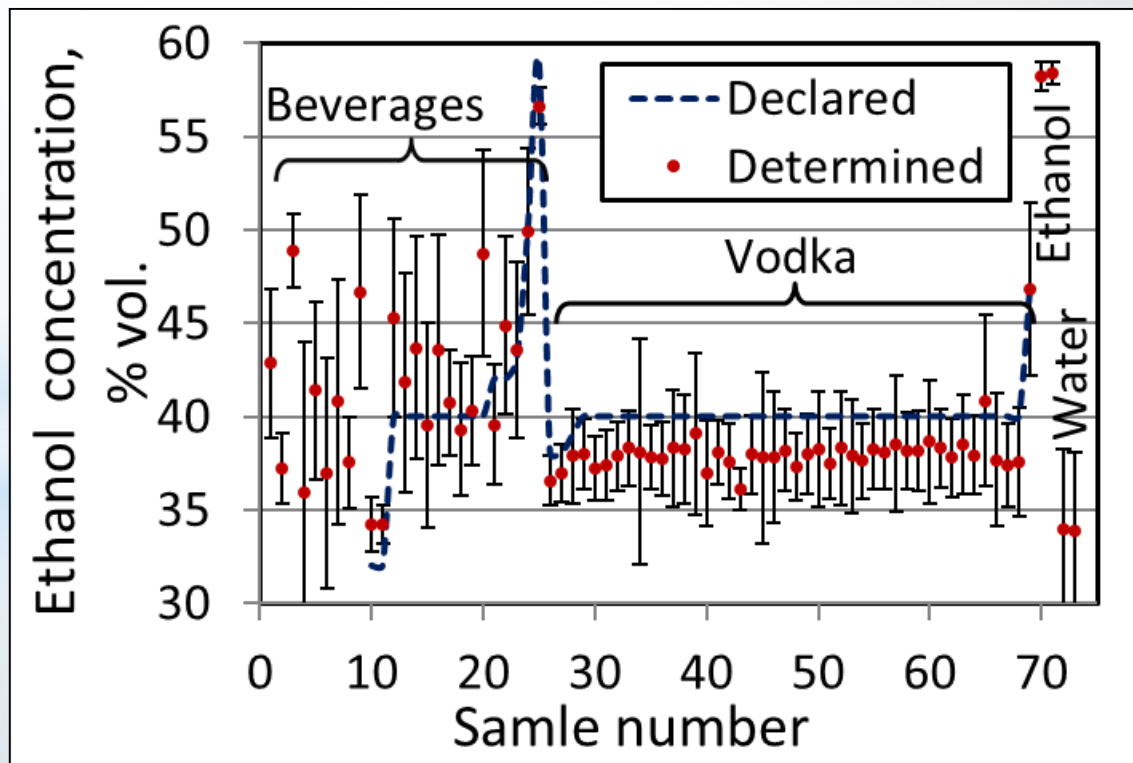


При использовании НС, обученных с добавлением шума, **результаты** определения концентраций оказались **близки к заявленным** производителями.

Задача “Спирты”. Растяжение/сжатие спектров.

Набор данных реальных алкогольных напитков

Сети, обученные с добавлением 20% Гауссова шума

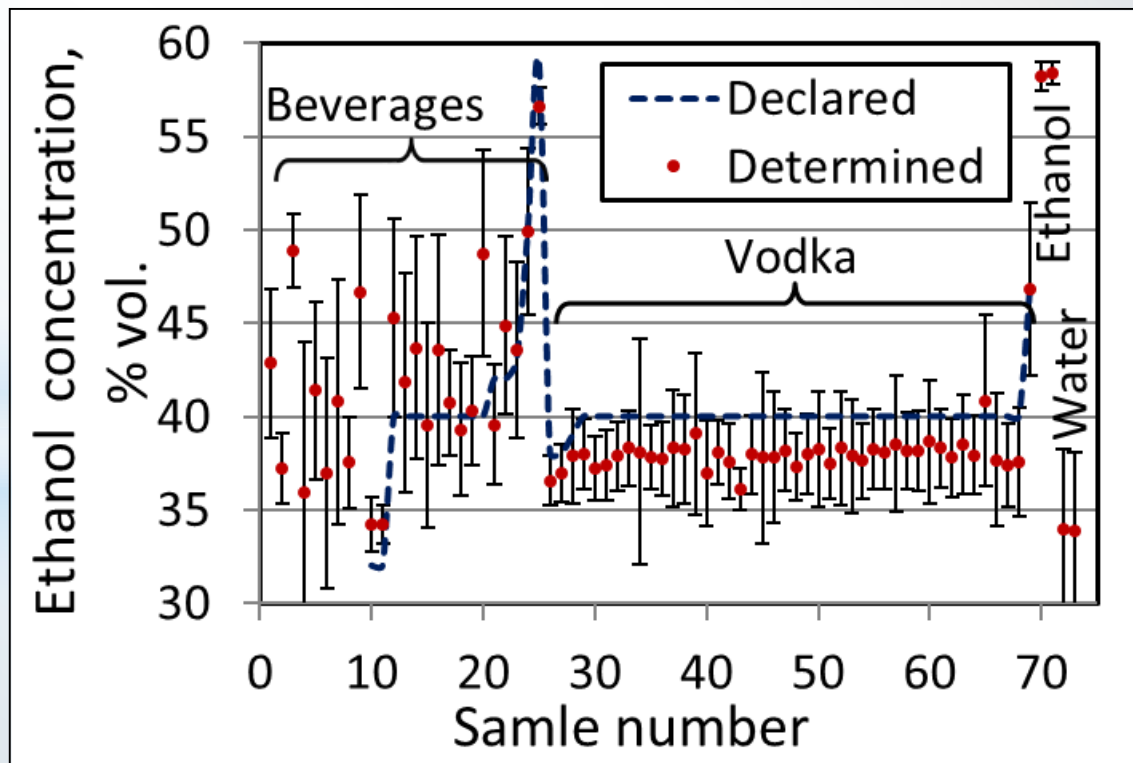


Наблюдаемая систематическая ошибка косвенно свидетельствует о наличии неучтенных искажений в наборе.

Задача “Спирты”. Растяжение/сжатие спектров.

Набор данных реальных алкогольных напитков

Сети, обученные с добавлением 20% Гауссова шума



Среднее отклонение концентрации составило 2,07% об.

Выводы

Выводы

- ❑ Показана принципиальная возможность применения методов машинного обучения для обработки данных обратных задач оптической спектроскопии.
- ❑ Описаны особенности обратных задач оптической спектроскопии и способы их учета.
- ❑ Продемонстрированы результаты использования методов повышения устойчивости нейросетевого решения к шумам в данных

Спасибо за внимание!