**Московский Государственный Университет имени М.В. Ломоносова**

**Физический факультет**

**Кафедра атомной физики, физики плазмы и микроэлектроники**

*Курсовая работа студента 2 курса Сухадольского А.Г.*

**ОПТИМИЗАЦИЯ ПРЕДОБРАБОТКИ ВХОДНЫХ ПРИЗНАКОВ ПРИ ПРОГНОЗИРОВАНИИ ГЕОМАГНИТНОГО ИНДЕКСА DST МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ**

**OPTIMIZATION OF PREPROCESSING OF INPUT FEATURES IN PREDICTING THE GEOMAGNETIC DST INDEX USING MACHINE LEARNING METHODS**

Научные руководители:

Зав. лаб., к.ф.-м.н. Доленко С. А.

Инженер Владимиров Р.Д.

Оглавление

[1. Введение. 3](#_Toc166745903)

[2. Цель и задачи работы 3](#_Toc166745904)

[3. Обзорная часть 4](#_Toc166745905)

[3.1. Предметная область 4](#_Toc166745906)

[3.2. Методы обработки данных 7](#_Toc166745907)

[4. Вычислительный эксперимент 9](#_Toc166745908)

[5. Результаты 10](#_Toc166745909)

[6. Выводы 14](#_Toc166745910)

[7. Список литературы 14](#_Toc166745911)

# Введение.

Магнитным полем Земли или геомагнитным полем называют поле, генерируемое внутриземными источниками. Оно испытывает возмущения, называемые геомагнитными вариациями, при взаимодействии его внешнего слоя, магнитосферы, с солнечным ветром - потоком плазмы, электронов и ионов, преимущественно протонов, из солнечной короны. [1]

Для оценки состояния магнитосферы Земли используют различные геомагнитные индексы, являющиеся количественной мерой геомагнитной активности и предназначенные для описания вариаций магнитного поля Земли. В данной работе взят за основу индекс шторм-тайм вариации (storm-time variation) Dst [2]который представляет собой максимальное отклонение от спокойного уровня на контрольной цепочке магнитных станций. Dst-индекс вычисляется один раз за часЭто один из наиболее часто используемых индексов при исследовании космической погоды и её влияния на магнитосферу Земли.

Геомагнитные вариации оказывают влияние на работу космических аппаратов, кроме того, сильные геомагнитные возмущения, называемые геомагнитными бурями, нарушают функционирование наземных технических устройств, например, энергосетей и радиосвязи, из-за кольцевых токов, возникающих в магнитосфере во время магнитных бурь, и сильных токов, наводимых в проводниках.

В связи с этим особую важность играет прогнозирование такого рода возмущений, которое, однако, затрудняется тем фактом, что магнитосфера Земли представляет собой многокомпонентную динамическую систему, открытую для внешних воздействий. В совокупности с чрезвычайной сложностью описания процессов в системе «Солнце - солнечный ветер - магнитосфера Земли» это делает построение хорошей физической модели очень непростой задачей, в связи с чем в данной работе используются методы машинного обучения (ММО).

1. Цель и задачи работы

Одним из важнейших этапов машинного обучения является предобработка данных, от которой зависит всё дальнейшее обучение. В частности, на качество прогнозирования большое влияние оказывает распределение данных, как для входных признаков, так и для выходного признака (целевой переменной, значение которой прогнозируется). К сожалению, распределение значений индекса Dst весьма далеко как от равномерного, так и от нормального.

Целью данной работы является проверка предположения о том, что качество прогнозирования Dst-индекса с помощью ММО может быть повышено путем предобработки его значений с помощью нелинейного преобразования.

Задачами, решаемыми для достижения данной цели, были:

1) Программная реализация обучения и применения искусственной нейронной сети (НС) типа многослойный персептрон (МСП).

2) Исследование результатов прогнозирования индекса Dst с горизонтом от 1 до 24 часов с помощью линейной регрессии и МСП.

3) Исследование результатов прогнозирования индекса Dst с использованием тех же ММО с предварительным нелинейным преобразованием его значений путём логарифмирования с предварительным смещением и их сравнение с результатами, полученными без преобразования.

1. Обзорная часть
   1. Предметная область

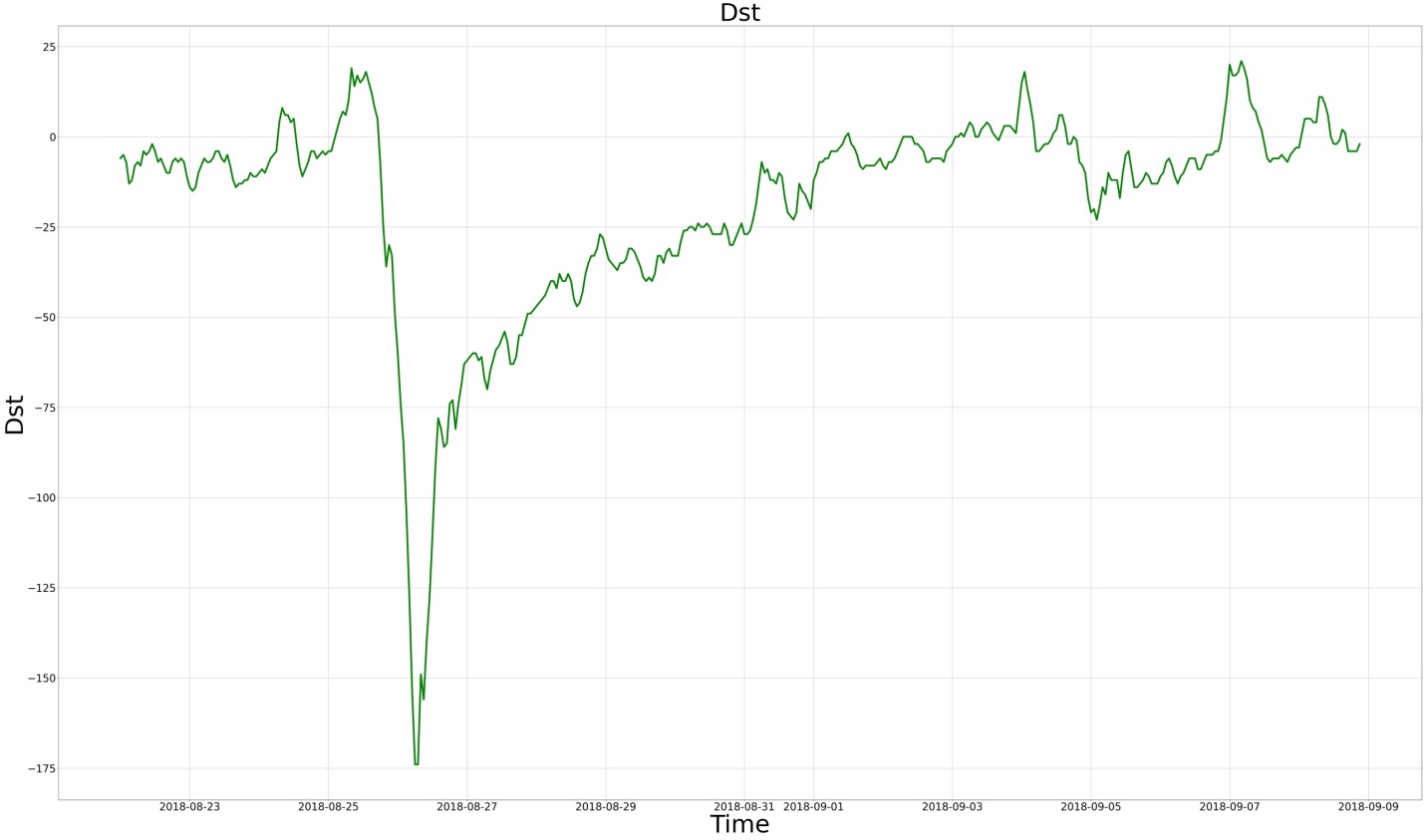
Рассмотрим механизм геомагнитных возмущений. Воздействие солнечного ветра на магнитосферу Земли вызывает поджатие магнитосферы в головной части и возрастание напряженности поля. Сжатие же создает электрически нейтральная плазма солнечного ветра. При столкновении облака плазмы с внешним магнитным полем Земли, отдельные протоны и электроны отражаются под действием силы, перпендикулярной к магнитному полю и направлению движения частиц. Далее, движение заряженных частиц естественным образом создает электрический ток, который в свою очередь создает собственное магнитное поле. Индуцированное поле уничтожает нормальное геомагнитное поле в облаке и усиливает нормальное поле в области между облаком и Землей. Облако продолжает продвигаться до тех пор, пока его энергия не израсходуется на сжатие геомагнитного поля. При этом возникшая в результате граница называется головной ударной волной. [3]

Сильные геомагнитные возмущения (геомагнитные бури) делятся на следующие характерные временные интервалы, называемые фазами бури: начальную фазу (может отсутствовать), главную фазу и фазу восстановления.

Буря часто начинается с резкого увеличения индекса Dst. При этом временной интервал между внезапным началом бури и моментом уменьшения поля называется начальной фазой.

Уменьшение поля может продолжаться несколько часов и называется главной фазой. Она характеризуется быстрым падением значения индекса Dst.

Затем начинается медленное восстановление, и индекс Dst возвращается к уровню, который был перед бурей, или во многих случаях к уровню немного ниже исходного.

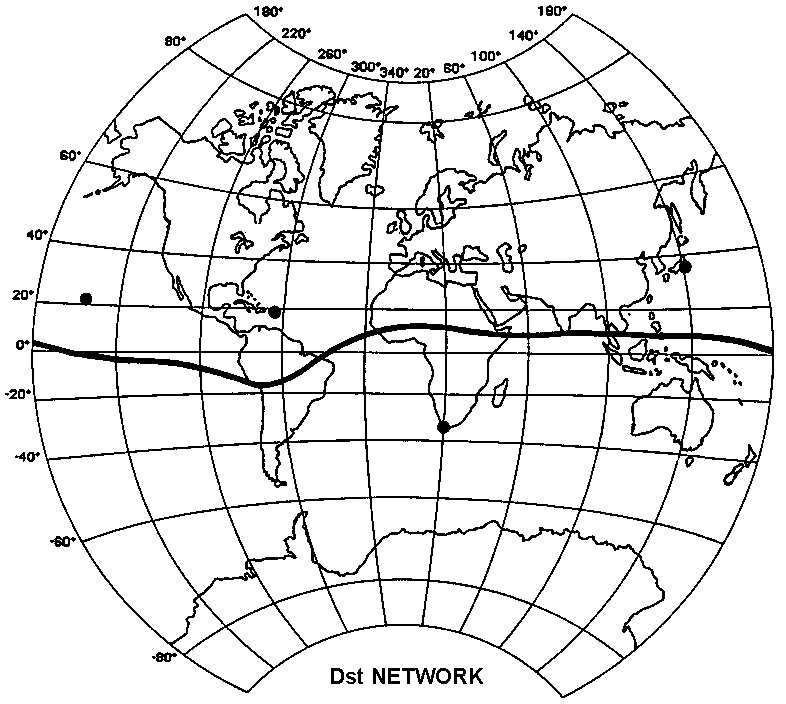


*Рис. 1. Характерная динамика индекса Dst во время геомагнитной бури*

Заметим, что бури хорошо выделяются в средних широтах, но в зоне полярных сияний и в более высоких широтах вариации носят другой характер из-за наличия авроральной электроструи. Также межпланетное поле может пересоединяться с геомагнитным полем и таким образом непосредственно влиять на магнитные вариации.

Хочется отметить, что уменьшение поля иногда может наблюдаться и без внезапного начала бури, а также протекать в различных иррегулярных формах. Однако в последние годы термин "магнитная буря" чаще употребляется в том случае, если имеются все характерные фазы: внезапное начало, главная фаза, фаза восстановления.

Сам Dst-индекс характеризует вариацию полей из нескольких источников: поля токов, текущих вдоль поверхности Земли (ими определяется положительная вариация начальной фазы), поля кольцевого тока и поля токов, текущих в ионосфере, наиболее значимые в полярных районах. Для получения индекса Dst используются четыре магнитные обсерватории: Херманус, Какиока, Гонолулу и Сан-Хуан. Эти обсерватории были выбраны по качеству наблюдений, а также по той причине, что их расположение достаточно удалено от авроральных и экваториальных электроджетов и максимально равномерно распределено по долготе. Контрольный уровень для Dst установлен таким образом, чтобы в пять самых тихих дней, обозначенных на международном уровне, индекс Dst в среднем был равен нулю [2].



*Рис. 2. Расположение магнитных обсерваторий.*

Основные возмущения в Dst носят характер понижения геомагнитного поля. Эти ослабления поля производятся главным образом экваториальной токовой системой в магнитосфере, обычно называемой кольцевым током. Ток нейтрального слоя, текущий через хвост магнитосферы, вносит небольшой вклад в уменьшение поля вблизи Земли. Положительные изменения Dst в основном вызваны сжатием магнитосферы из-за увеличения давления солнечного ветра [3].

Значение индекса Dst зависит от параметров плазмы солнечного ветра – скорости и плотности, параметров межпланетного магнитного поля – модуля вектора магнитной индукции и его z-компоненты (в системе координат GSM), а также от положения Земли относительно Солнца, вращения Земли, и от предыстории всех данных значений. Чтобы исключить зависимость параметров солнечного ветра и параметров межпланетного магнитного поля от положения измерительного прибора в пространстве, необходимо, чтобы космический аппарат, который проводит измерения, всегда находился на одном и том же постоянном расстоянии от Земли и от Солнца. Для выполнения данного условия подходит точка Лагранжа – точка гравитационного равновесия, которая находится в 1,5 миллиона километров от Земли. В данной точке (на орбитах малого радиуса вокруг неё) возможно обеспечить длительное нахождение аппарата. При этом сигнал от него доходит до Земли за 5 секунд, тогда как солнечный ветер проходит данное расстояние за время около 40 минут.

Данные для настоящей работы были взяты из Центра Анализа Данных для Геомагнитизма и Космического магнетизма Университета Киото (Data Analysis Center for Geomagnetism and Space Magnetism Graduate School of Science, Kyoto University) [4].

3.2. Методы обработки данных

В данной работе, как уже было сказано выше, были применены методы машинного обучения - класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение (подстройка параметров модели) за счёт использования результатов решения множества сходных задач. Конкретно в данной работе использовалась линейная регрессия с различными способами регуляризации, а также нейронная сеть вида многослойный персептрон с единственным скрытым слоем.

Регрессия — способ построить на основании семейства функций модель, которая минимизирует функцию потерь. Последняя характеризует то, насколько сильно модель отклоняется от желаемых значений в заданных точках. В данной работе использовалась линейная регрессия, т.е. семейство функций, из которых мы выбираем, представляет собой линейную комбинацию наперед заданных базисных функций

Ридж-регрессия (ridge regression, гребневая регрессия) использует регуляризацию по метрике L2 - вводит квадратичный штраф (penalty) за слишком большие коэффициенты, минимизируя следующую функцию:

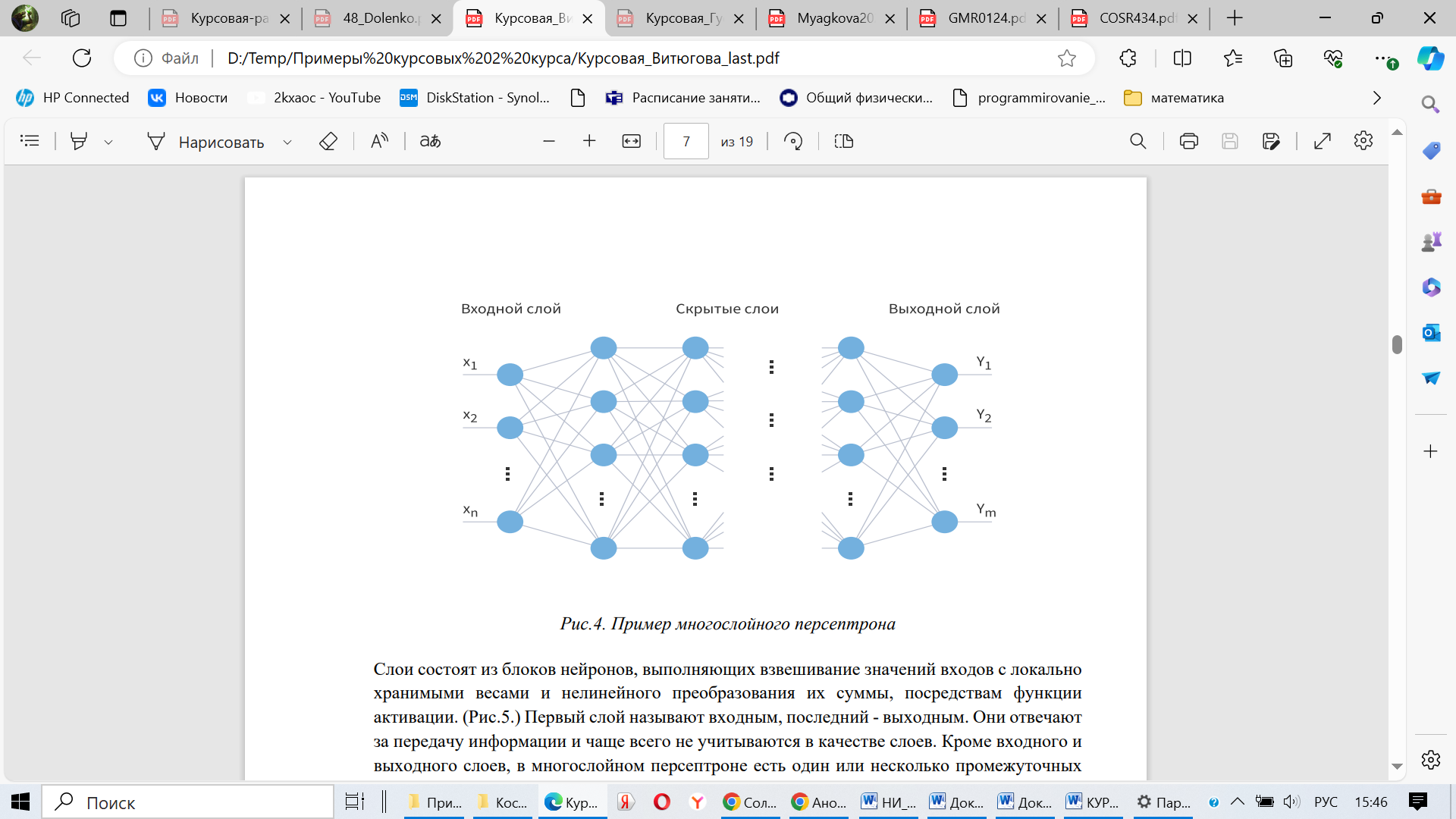
Здесь y – целевая функция, x – признаки, w - веса, – параметр.

Регрессия лассо (Lasso) использует для регуляризации сумму весов по модулю, то есть L1-норму:

Здесь N - количество примеров, остальное аналогично ридж-регрессии.

Отличие L1 и L2 регуляризации заключается в том, что гребневая регрессия при возведении коэффициентов в квадрат сильнее штрафует большие коэффициенты, а лассо регрессия штрафует все коэффициенты одинаково. Также L1 регуляризация в отличие от L2 может полностью обнулить часть коэффициентов и, таким образом, исключить соответствующие признаки из модели.

Однако основной моделью для предсказания индекса Dst в данной работе являлась искусственная нейронная сеть - модель в машинном обучении, построенная на принципе биологических нейронных сетей.



*Рис. 3. Пример структуры многослойного персептрона.*

Сеть включает в себя входной слой, куда подаются входные признаки. Далее они с различными коэффициентами, называемыми весами, передаются в первый и последующие (если есть) скрытые слои. В конце находится выходной слой, генерирующий итоговый результат работы модели.

В процессе обучения моделей машинного обучения используют два набора данных - тренировочный и тестовый. Тренировочным называется набор данных, который используется для обучения (подстройки параметров). Для проверки качества работы сети используются наборы, составленные из независимых данных, которые не использовались в процессе обучения - тестовые.

1. Вычислительный эксперимент

Для обучения моделей за тренировочный набор в данной работе были взяты данные с 1997 по 2016 годы, 157740 примеров. В качестве входных данных моделей использовались значения следующих физических величин:

* Скорость солнечного ветра в точке Лагранжа L1 v (км/с)
* Плотность протонов в солнечном ветре в точке Лагранжа L1(
* Векторные компоненты межпланетного магнитного поля: , , (соответствующие компоненты магнитного поля) и (модуль вектора магнитной индукции) (нТ)

Было произведено так называемое погружение многомерного временного ряда, состоящего из перечисленных величин, на 24 часа, т.е. массив входных признаков был дополнен значениями каждой из этих величин за предыдущие 24 часа. Погружение производилось как с обычными данными, так и с предварительно логарифмированными.

В рамках проведения вычислительного эксперимента были использованы готовые реализации линейной регрессии с линейно-алгебраическим вычислением коэффициентов, а также с L1 и L2 регуляризацией из библиотеки scikit-learn [5], а также реализация многослойного персептрона из той же библиотеки. Использовался персептрон с одним скрытым слоем из 64 нейронов. На вход персептрона поступает 148 параметров. В качестве функции активации как в скрытом, так и в выходном слое используется тангенсальная функция. Решение идет стохастический градиентный спусковой метод Адам (Adaptive Moment Estimation). Гиперпараметры для регуляризаций L1 и L2 и для персептрона были подобраны методом grid search –поиском по сетке гиперпараметров, с выбранной метрикой .

Логарифмирование значений индекса Dst осуществлялось с использованием следующей формулы:

Dstlog = lg (100-Dst)

В случае использования логарифмирования оно применялось как к входным признакам, содержащим значение индекса Dst со всеми задержками, так и к выходному признаку (прогнозируемой переменной). При этом перед вычислением критериев качества прогнозирования осуществлялось обратное преобразование прогноза из логарифмированного представления в исходное.

Осуществлялось прогнозирование значения индекса Dst с горизонтами от 1 до 24 часов с использованием вышеперечисленных методов, а также с помощью тривиальной инерционной модели, прогноз которой равен последнему известному значению прогнозируемой величины.

В качестве критериев оценки качества прогноза были выбраны следующие метрики (вычислялись на тестовом наборе данных):

1) СКО (RMSE) - средняя квадратичная ошибка (Root mean squared error)

2) Коэффициент детерминации :

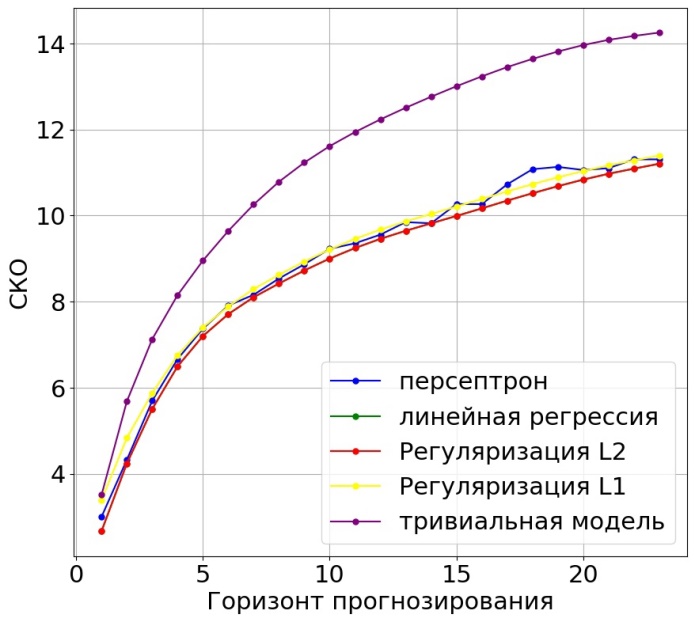
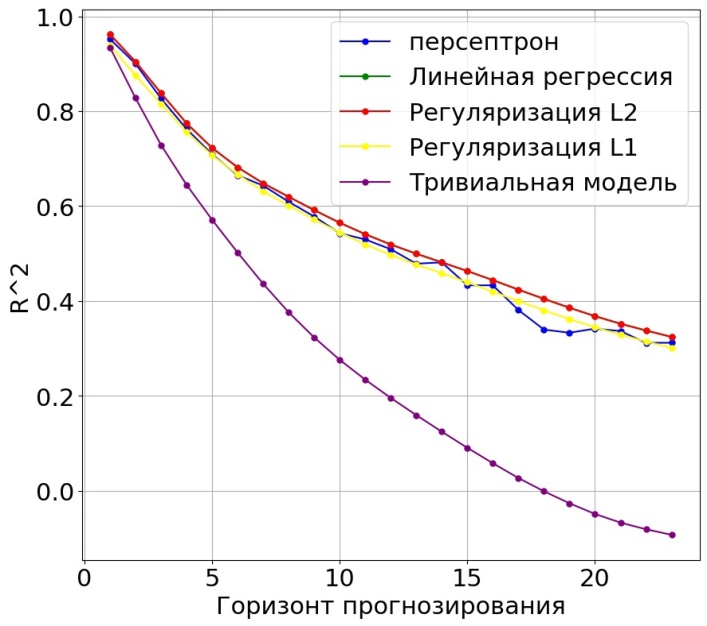
Здесь *n* – число примеров в выборке данных, на которой производится оценка, *xi* – ответ модели для *i*-го примера выборки, *x*ист *i* – истинное значение прогнозируемой переменной для *i*-го примера выборки, – среднее по выборке истинное значение прогнозируемой переменной.

1. Результаты

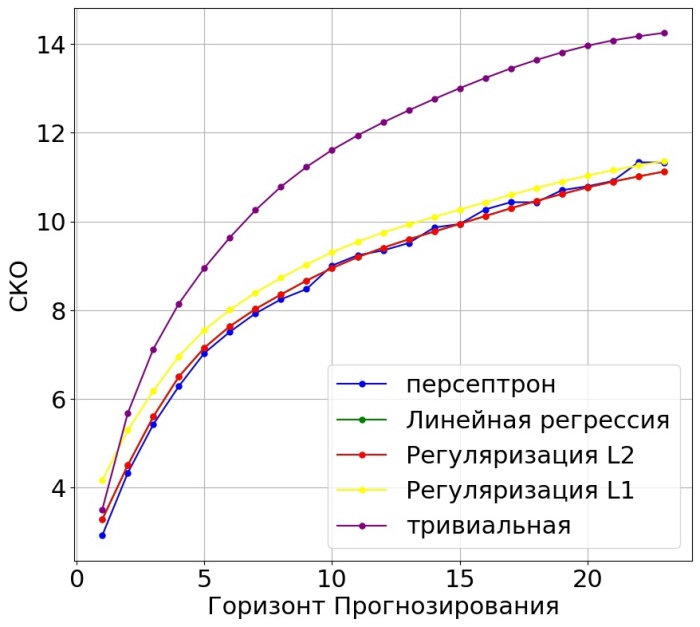
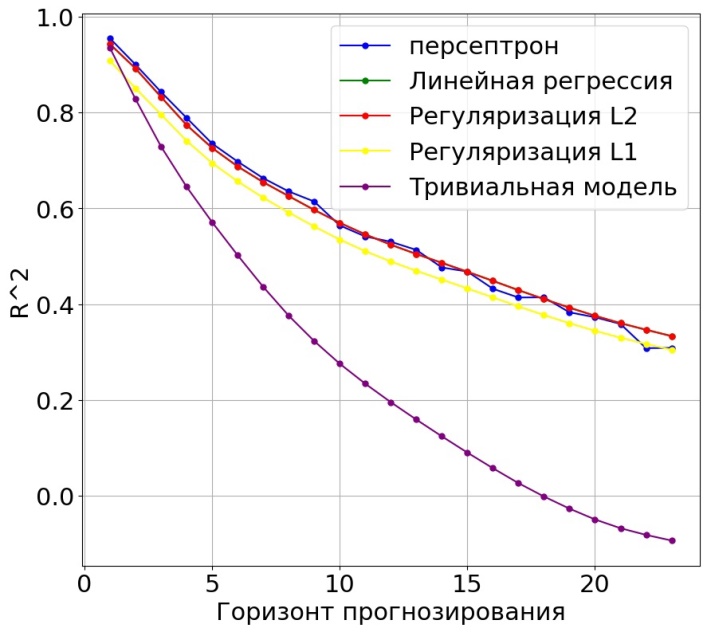
Рассмотрим полученные в данной работе результаты.

На Рис. 4 приведены зависимости критериев качества прогнозирования от горизонта прогноза для всех построенных моделей, при использовании индекса Dst в исходном представлении. На Рис. 5 приведены те же зависимости при использовании индекса Dst в логарифмированном представлении. При этом хорошо заметно монотонное убывание качества прогноза всех моделей с увеличением горизонта. Минимально приемлемым можно считать качество прогноза при горизонте не более 12 часов.

Наилучшие результаты прогнозирования показал многослойный персептрон и регуляризация L2, что может быть объяснено тем, что искомое отображение является нелинейным. Результатом этого является более низкое качество прогнозирования, показанное всеми более линейными моделями.



*Рис. 4. Зависимость качества прогнозирования от горизонта прогноза. Индекс Dst в исходном представлении.*

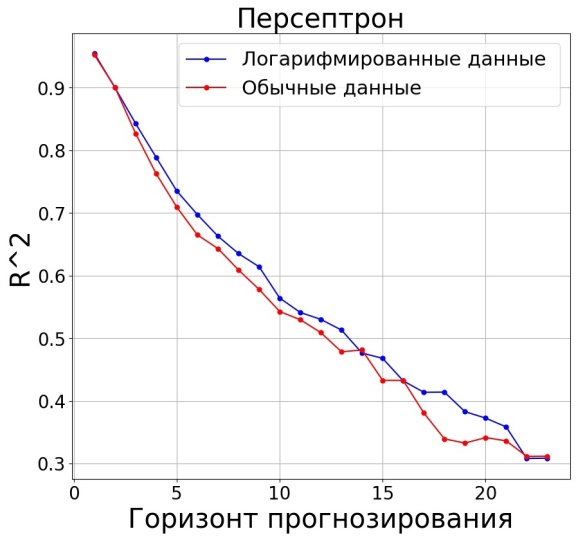
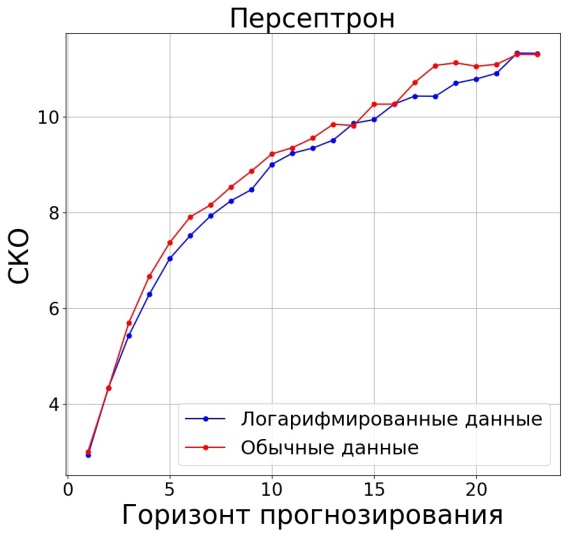


*Рис. 5. Зависимость качества прогнозирования от горизонта прогноза. Индекс Dst в логарифмическом представлении.*

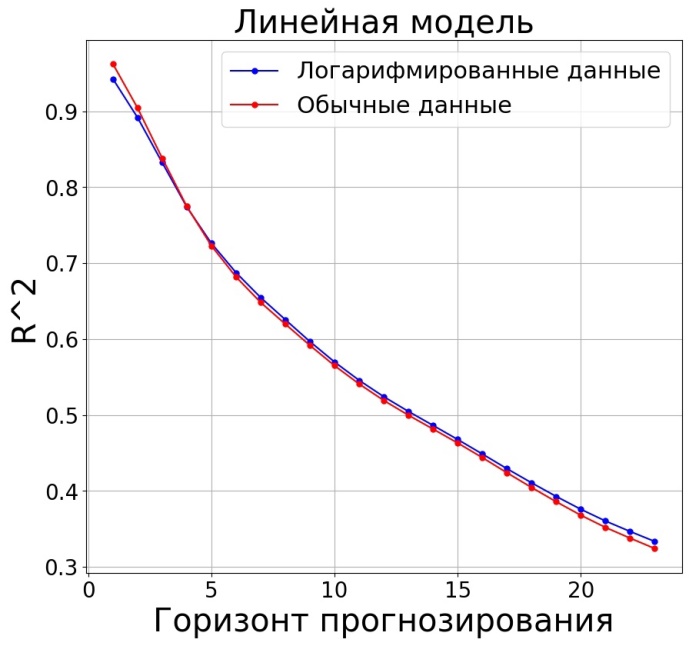
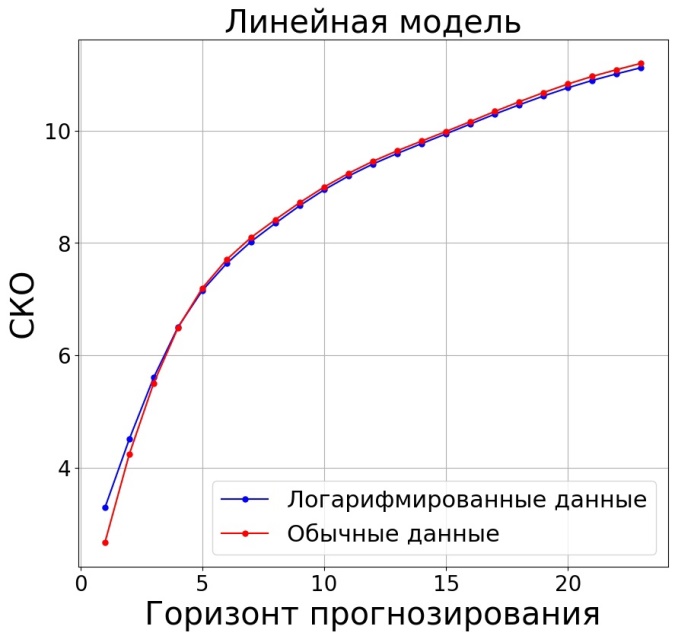
Так как полученные зависимости для исходного и для логарифмированного представлений индекса Dst оказались сходными, на Рис. 5 приведено сравнение полученных зависимостей для обоих представлений для многослойного персептрона.

Видно, что логарифмирование позволяет повысить качество прогноза только для значений горизонта прогнозирования 2,5 часов и более. По-видимому, при меньших значениях горизонта искомая зависимость является менее нелинейной, и логарифмирование как нелинейное преобразование усложняет её построение.

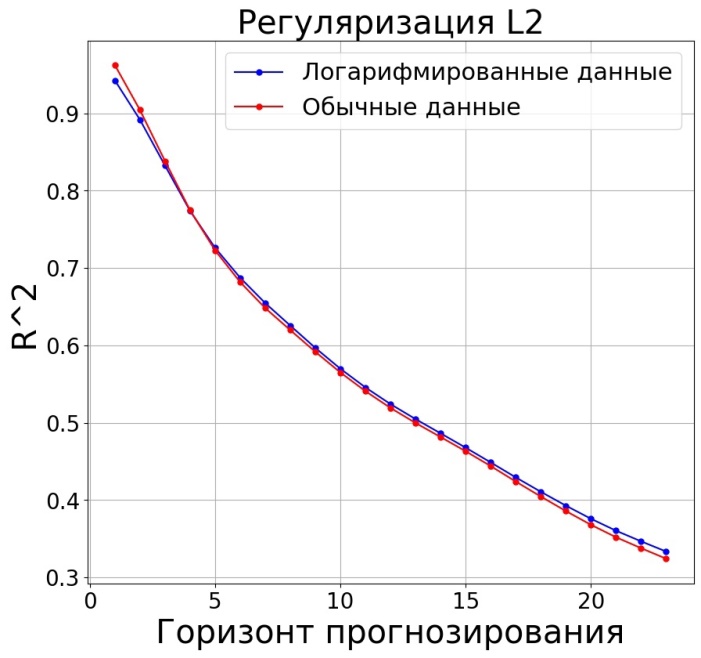
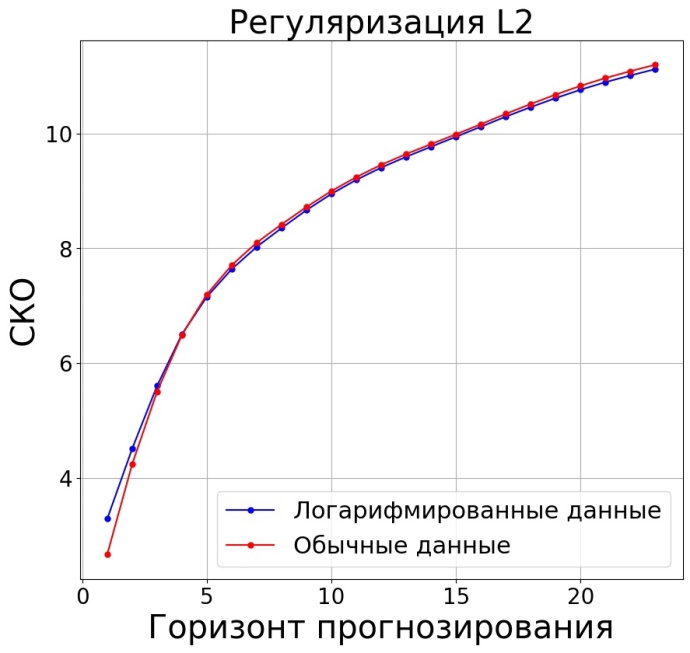
На Рис. 6-9 приведено аналогичное сравнение для других моделей – различных видов линейной регрессии. Видно, что качественно картина не меняется – логарифмическое преобразование значений индекса Dst позволяет повысить качество прогнозирования только при значениях горизонта прогноза, превышающих определённый порог. Сами пороговые значения для различных видов моделей могут отличаться.



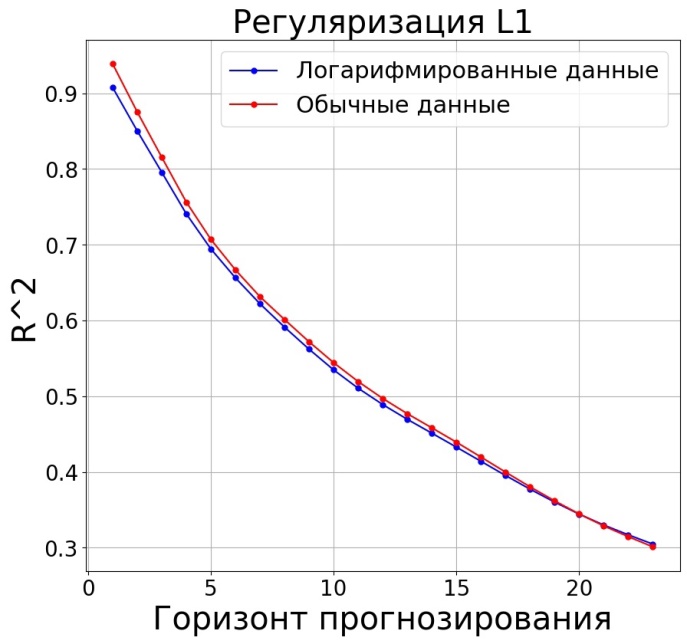
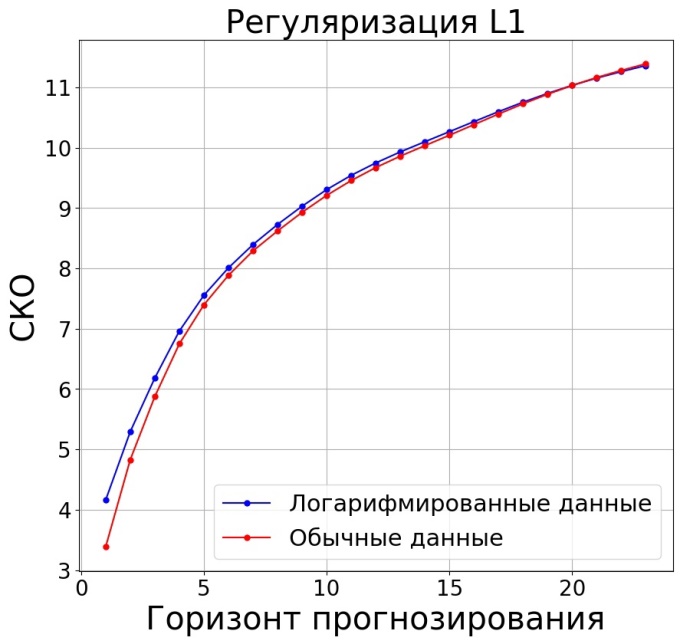
*Рис. 6. Сравнение зависимостей качества прогнозирования многослойным персептроном от горизонта прогноза для исходного и логарифмированного представления индекса Dst.*



*Рис. 7. Сравнение зависимостей качества прогнозирования с помощью линейной регрессии без регуляризации от горизонта прогноза для исходного и логарифмированного представления индекса Dst.*



*Рис. 8. Сравнение зависимостей качества прогнозирования с помощью линейной регрессии с регуляризацией L2 (гребневая регрессия) от горизонта прогноза для исходного и логарифмированного представления индекса Dst.*



*Рис. 9. Сравнение зависимостей качества прогнозирования с помощью линейной регрессии с регуляризацией L1 (лассо регрессия) от горизонта прогноза для исходного и логарифмированного представления индекса Dst.*

1. Выводы

По результатам настоящей курсовой работы могут быть сделаны следующие выводы.

1) Для всех построенных моделей качество прогнозирования монотонно убывает с увеличением горизонта прогноза. При этом минимально приемлемым можно считать качество прогнозирования при горизонте прогноза не более 12 часов.

2) Построенные модели могут быть ранжированы по качеству прогнозирования в следующем порядке: многослойный персептрон, гребневая регрессия, линейная регрессия без регуляризации, лассо регрессия, тривиальная модель.

3) Использование логарифмированных данных повышает точность прогнозирования при горизонте прогноза 2,5 часов и более. При меньших величинах горизонта качество прогнозирования ухудшается на линейных моделях и почти не меняется на персептроне, что может быть связано с низкой нелинейностью требуемого отображения, которое более эффективно аппроксимируется в исходном не логарифмированном представлении.

1. Список литературы

[1]. Akasofu S.-I., Chapman S. Solar-Terrestrial Physics. Oxford: Clarendon Press, 889 p. 1972.

[2]. Toyohisa Kamei, Masahi Sugiura. On Dst index (description in the IAGA Bulletin No 40) (https://wdc.kugi.kyoto-u.ac.jp/dstdir/dst2/onDstindex.html)

[3]. Сервер «СиЗиФ» (Солнечно-Земная физика, <http://www.kosmofizika.ru/>)

[4]. Всемирный центр данных в Киото (Geomagnetic Data Service.

http://wdc.kugi.kyoto-u.ac.jp/wdc/Sec3.html)

[5]. Библиотека Scikit-learn (https://scikit-learn.org/stable/index.html)