УДК 504.064.37

**М.Ю. Катаев, С.Г. Катаев, А.Г. Андреев, С.А. Базелюк, А.К. Лукьянов**

# Непараметрические математические методы восстановления общего содержания $CO_2$ из данных спутникового мониторинга

Приводится сравнение непараметрических подходов при решении задачи восстановления общего содержания углекислого газа по модельным данным измерений спутниковым прибором GOSAT. Приводятся результаты тестовых расчетов спектров отраженного от поверхности солнечного излучения в ближней ИК-области спектра.

**Ключевые слова:** отраженное от поверхности солнечное излучение, спутниковый спектрометр, нейронные сети, случайные деревья, эмпирические ортогональные функции.

Одной из самых волнующих человечество проблем является потепление климата. Важнейшим фактором потепления называют увеличение концентрации парниковых газов. В связи с этим возникает задача круглогодичного мониторинга концентрации этих газов во всей земной атмосфере. Методы спутникового мониторинга занимают одно из ведущих мест в этом направлении, благодаря возможности оперативно получать данные с высоким пространственным и временным разрешением одновременно в широком спектральном диапазоне. В 2009 г. запущен спутник IBUKI с прибором GOSAT (Японское космическое агентство [http://www.jaxa.jp]) на борту, который предназначен для мониторинга концентрации $CO_2$ и $CH_4$. Японская сторона [www.gosat.nies.go.jp] разработала собственные методики обработки спутниковых данных и позволила мировому научному сообществу присоединиться к разработке методик, отвечающих условиям точности, стабильности, скорости и устойчивости. Нами предлагается для целей обработки данных спутникового мониторинга общего содержания $CO_2$ применять непараметрические подходы решения обратной задачи.

**Постановка задачи**

Взаимодействие солнечного излучения с атмосферой приводит к рассеянию и поглощению фотонов солнечного излучения и количественно определяется свойствами газового состава и типами аэрозоля [1]. Излучение, которое было отражено от поверхности или облаков, зависит от характера подстилающей поверхности, отражающих свойств и температуры поверхности. Какая-то часть солнечного излучения, достигшая спутникового прибора, зависит от поглощающих свойств газового состава и таким образом может быть использована для определения содержания газового состава атмосферы. Процессы, сопровождающие прохождение солнечного света в системе «Земля + Атмосфера», схематически показаны на рис. 1. Здесь показаны траектории солнечных лучей: рассеянных (однократно или многократно) в атмосфере, отраженных от поверхности Земли, от облаков и зарегистрированных спутником.

Пусть мы имеем набор измерений $\mathbf{Y}$ отраженного от поверхности солнечного излучения, который содержит в себе информацию об искомом параметре (общем содержании $CO_2$ и $CH_4$ на оптической трассе формирования сигнала). Между сигналом и искомым параметром существует некоторая функциональная связь в виде выражения

$$\mathbf{Y}(i, j) = \mathbf{F}(\mathbf{x}(j)), \qquad (1)$$

здесь $\mathbf{Y}$ – измеряемый сигнал; $\mathbf{F}$ – функционал, описывающий трансформацию излучения Солнца в ближней ИК-области спектра по трассе измерений, и $\mathbf{x}$ – искомый параметр; $i = 1\ldots m$ – число спектральных каналов; $j = 1\ldots n$ – число измерений в течение определенного времени (например, год, в течение которого искомый параметр испытывает изменение, связанное с различными физическими процессами).

Для обработки спутниковых сигналов, с целью восстановления общего содержания $CO_2$, как правило, применяется метод оптимального оценивания [2], который относится к классу параметрических. В таких подходах требуется четкое соответствие модели измерений и измеряемого сигнала. Известно, что между моделью и реальным сигналом может быть отличие, которое в сумме с ошибками измерений является источником ошибок восстановления неизвестных параметров модели.

Нами предлагается применить непараметрические подходы, которые из закономерностей изменения сигнала, в зависимости от вариации искомого параметра, позволяют построить модель, кото-

рая будет для данной величины сигнала и условий измерения выдавать значение искомого параметра. В статье нами рассматриваются три известных подхода: метод эмпирических ортогональных функций, метод нейронных сетей и метод случайных деревьев.

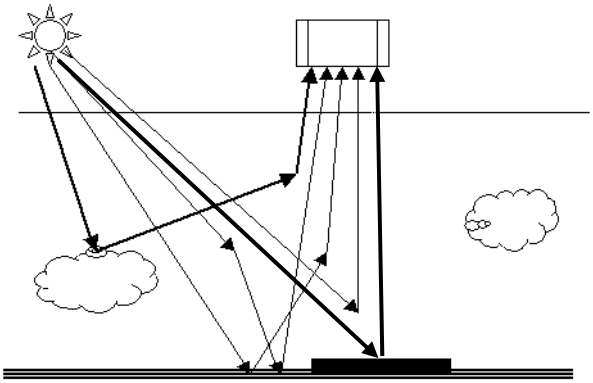

Рис. 1. Схема траекторий солнечных лучей в системе «Земля + Атмосфера»

**Метод эмпирических ортогональных функций (ЭОФ)**

Принцип метода ЭОФ широко представлен в литературе, в основном для анализа рядов наблюдений, сжатия информации, выявления закономерностей проявления физических процессов во времени и пространстве [1, 3]. Гораздо реже этот подход используется при решении обратных задач. Первым шагом, в построении регрессионной модели на основе ЭОФ является построение корреляционной матрицы:

$$\mathbf{C}(i,l)=\sum_{j=1}^{n}(\mathbf{Y}(i,j)-\overline{\mathbf{Y}}(j))(\mathbf{Y}(l,j)-\overline{\mathbf{Y}}(l)), \quad (2)$$

где $\overline{Y}$ – осредненный сигнал для всех измерений $n$ ($\mathbf{C}$ имеет размерность [$m\times m$]).

Далее корреляционная матрица раскладывается по собственным векторам и значениям:

$$\sum_{j=1}^{m}\mathbf{C}(i,j)\mathbf{E}(j,k)=\mathbf{\Lambda}(k)\mathbf{E}(i,k), \quad (3)$$

здесь $\mathbf{C}$ – корреляционная матрица; $\mathbf{E}$ – матрица собственных векторов (имеет размерность [$m\times m$]) и $\mathbf{\Lambda}$ – вектор собственных значений (имеет размерность [$m$]).

На основе собственных векторов можно построить главные компоненты (или ЭОФ):

$$\mathbf{G}(i,k)=\sum_{j=1}^{m}\mathbf{E}(j,i)(\mathbf{Y}(j,k)-\overline{\mathbf{Y}}(j)), \quad (4)$$

где $k = 1...n$ (число измерений), $i = 1...p$, ($p \le m$).

Для решения обратной задачи необходимо учитывать тот факт, что сигналы связаны с общим содержанием газов, поглощающих солнечное излучение в выбранном канале длин волн. Тогда линейная регрессия искомой величины, с учетом главных компонент, может быть представлена следующим образом:

$$\mathbf{X}(i)=\sum_{j=1}^{p}\mathbf{A}(j)\mathbf{G}(i,j)+\overline{\mathbf{X}}, \quad (5)$$

здесь $i = 1...n$ (число измерений), $j = 1...p...m$.

Решение системы линейных алгебраических уравнений (5) позволяет найти коэффициенты $\mathbf{A}$ и тем самым построить модель для обработки данных измерений в виде

$$\mathbf{X}_R=\sum_{j=1}^{p}\mathbf{A}(j)\sum_{l=1}^{p}\mathbf{E}(l,j)(\mathbf{Y}(l)-\overline{\mathbf{Y}})+\overline{\mathbf{X}}. \quad (6)$$

Выражение (6) позволяет получить искомое решение для выбранного спектра и получить общее содержание для данной географической точки (для которой получен сигнал). В нашем случае $\mathbf{Y}$ – это спектр, измеряемый спутником в ближней ИК-области спектра, а $\mathbf{X_R}$ – общее содержание $CO_2$.

**Метод нейронный сетей**

Нейронная сеть (НС) – вычислительная схема, построенная из однородных вычислительных элементов (нейронов), соединенных между собой определенными связями [4, 5]. Как правило, передаточные функции всех нейронов в нейронной сети фиксированы, а веса являются параметрами нейронной сети и могут изменяться. Некоторые входы нейронов помечены как внешние входы нейронной сети, а некоторые выходы – как внешние выходы нейронной сети. Подавая любые числа на входы нейронной сети, мы получаем какой-то набор чисел на выходах нейронной сети. Таким образом, работа нейронной сети состоит в преобразовании входного вектора в выходной вектор, причем это преобразование задается весами нейронной сети. Практически любую задачу можно свести к задаче, решаемой нейронной сетью.

Нами построена модель нейронной сети для восстановления общего содержания $CO_2$, обобщенная схема которой представлена на рис. 2 [6].

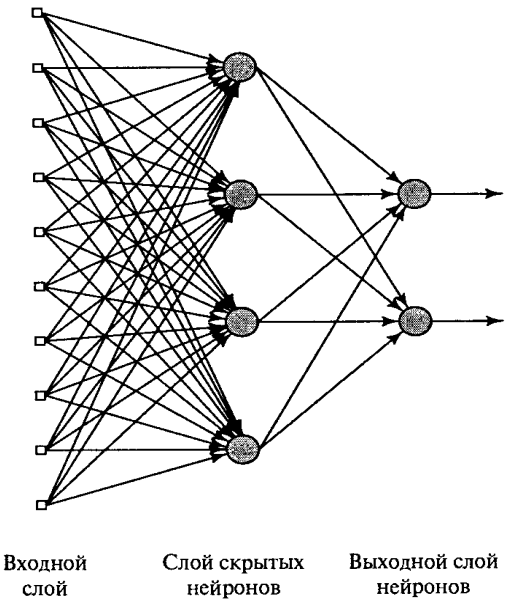


Рис. 2. Схематическое представление нейронной сети (многослойный персептрон)

Одним из самых важных свойств нейронной сети является её способность обучаться. В процессе обучения параметры нейронной сети интерактивно настраиваются так, чтобы впоследствии она смогла решать поставленную задачу автоматически. Обучение считается законченным, когда правильно выполняет преобразование на тестовых примерах и дальнейшее обучение не вызывает значительного изменения настраиваемых параметров. Далее, в процессе работы, сеть выполняет преобразование ранее неизвестных ей данных на основе сформированной в процессе обучения нелинейной модели процесса.

Для обучения сети был выбран метод Левенберга–Марквардта как наиболее быстрый и устойчивый из стандартных методов, использующих производные как первого, так и второго порядка. Также для обучения используется процедура кроссвалидации для предотвращения явления переобучения.

Для формирования архитектуры сети был разработан эвристический алгоритм [6], заключающийся в следующем:

1. Выбираем некоторое начальное количество длин волн, подаваемых на вход сети, количество слоёв и нейронов в каждом из слоёв. Задаём значения step_wave – шаг увеличения количества длин вон, step_neuron – шаг увеличения количества нейронов в одном из слоёв, допустимое количество итераций.
2. На каждом этапе:

2.1. Обучаем сеть 1 с увеличенным на величину step_wave количеством длин волн, по сравнению с сетью, полученной на предыдущем этапе.

2.2. Обучаем $N$ сетей, где $N$ – количество промежуточных слоёв сети, полученной на предыдущем этапе, в каждой из $N$ сетей в одном из слоёв количество нейронов увеличено на величину step_neuron.

2.3. Обучаем сеть 1, где количество промежуточных слоёв увеличено и в новом слое step_neuron нейронов.

Итого получаем $N + 2$ сети, выбираем из этих такую сеть, ошибка которой на тестовой выборке минимальна, сохраняем её как полученную на данном этапе.

3. Проверяем критерий останова:

3.1. Если превышено количество допустимых итераций, то останов, в противном случае перейти на шаг 3.2.

3.2. Если ошибка сети меньше 0,1%, то останов, в противном случае перейти на шаг 3.3.

3.3. Если ошибка сети полученной на данном этапе отличается от ошибки сети, полученной на предыдущем этапе, меньше чем на 0,0001, то останов, в противном случае перейти на шаг 2.

**Метод случайных деревьев**

Подход случайных деревьев (СД) к решению задач регрессии и классификации был разработан сравнительно недавно [7, 8]. Его идея очень проста и легко реализуется на практике. При решении задачи восстановления общего содержания газов по спутниковым измерениям, нами этот подход применяется впервые.

Кратко идею подхода можно описать так: пусть нам задано некоторое обучающее множество **T,** содержащее объекты (примеры), каждый из которых характеризуется m атрибутами, причем один из них указывает на принадлежность объекта к определенному классу.

Пусть через {**C**1**, C**2**, ... C***k***}** обозначены классы, тогда существуют 3 ситуации:

1) множество **T** содержит один или более примеров, относящихся только к одному классу **C***k*. Тогда дерево решений для **T** – это лист, определяющий класс **C***k*;

2) множество **T** не содержит ни одного примера, т.е. пустое множество. Тогда это снова лист, и класс, ассоциированный с листом, выбирается из другого множества, отличного от множества **T**, скажем, из множества, ассоциированного с родителем;

3) множество **T** содержит примеры, относящиеся к разным классам. В этом случае следует разбить множество **T** на некоторые подмножества. Для этого выбирается один из признаков, имеющий два и более отличных друг от друга значений {**O**1**, O**2**, ... O***n***}**. Множество **T** разбивается на подмножества **{T**1**, T**2**, ... T***n***}**, где каждое подмножество **T***i* содержит все примеры, имеющие значение **O***i* для выбранного признака. Эта процедура будет рекурсивно продолжаться до тех пор, пока конечное множество не будет состоять из примеров, относящихся к одному и тому же классу.

Нами реализован алгоритм [6] на основе оригинальных предложений автора [http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm].

**Результаты тестирования алгоритмов**

Специфика задачи восстановления концентрации газа по спектрам отраженного солнечного излучения состоит в том, чтобы найти такие участки спектра, где поглощение этим газом максимально, а поглощение остальными газами атмосферы минимально. Для $CO_2$ такой участок найден, он находится в диапазоне частот 6180–6260 $см^{-1}$, что проиллюстрировано на рис. 3.

Спутниковый сигнал рассчитывается на основе программы [9] и представляет собой отраженное от поверхности Земли солнечное излучение для территории Западной Сибири. Для моделирования условий, приближенных к реальности (изменение параметров атмосферы учитывалось на основе базы NCEP [http://www.ncep.noaa.gov/]), рассчитывался спектр для нескольких углов склонения Солнца (10, 30, 50 и 70°). Всего спектров 1460 из расчета 4 измерения в день. Спектр содержит 1000 длин волн. Из них было выбрано 400 из условия максимального изменения сигнала (по правилу наибольшей вариации за год), а именно:

$$\max(\mathbf{Y}_{ij}) - \min(\mathbf{Y}_{ij}) = \Delta(j),$$

где $\mathbf{Y}_{ij}$ – величина сигнала $i$-го измерения на длине волны $j$; $\Delta(j)$ – вариация сигнала за год на $j$-й длине волны.

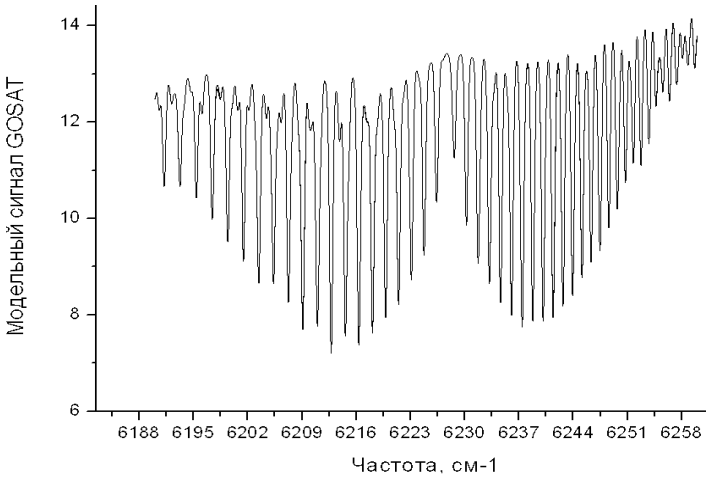


Рис. 3. Модельный сигнал прибора GOSAT для диапазона полосы поглощения $CO_2$

Далее полученные массивы данных разбивались на три выборки: 1 – одно измерение из каждого дня (365 измерений для одного угла склонения Солнца) и соответствующие им значения общей концентрации $CO_2$ – обучающая выборка; 2 – одно измерение из каждого дня, не совпадающее с первой выборкой (365 измерений для одного угла склонения Солнца) и соответствующие им значения общей концентрации $CO_2$ – валидационная выборка и 3 – оставшиеся измерения (730 измерений для одного угла склонения Солнца) и соответствующие им значения общей концентрации $CO_2$ –

тестовая выборка. К каждой выборке присоединяем вектор, элементами которого являются значения угла в радианах для данного измерения.

Для решения обратной задачи восстановления общего содержания $CO_2$ были разработаны программные коды, на основе описанных выше подходов:

1) ЭОФ – формулы (2)–(6);

2) НС – была выбрана нейронная сеть типа многослойный персептрон, метод Левенберга–Макрквардта для обучения сети, эвристический метод формирования топологии (количество длин волн, подаваемых на вход сети, количество слоёв и нейронов в каждом из слоёв);

3) СД – выбран авторский алгоритм [7].

Результаты решения обратной задачи восстановления общего содержания $CO_2$ приведены на рис. 4. Для определения работоспособности алгоритмов нами вычислялись статистические характеристики восстановленных значений общего содержания $CO_2$ и результаты приведены в таблице.

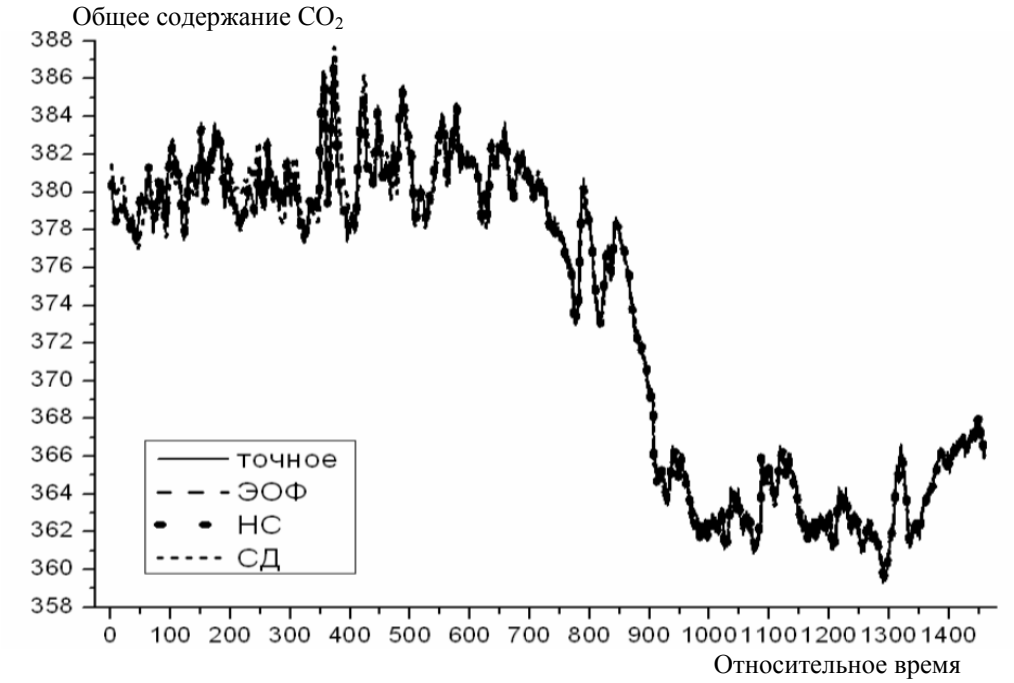


Рис. 4. Сравнение восстановленных значений общего содержания $CO_2$ (шкала абсцисс) из модельных данных измерений спутниковым прибором GOSAT для трех непараметрических методов: ЭОФ, НС и СД

**Максимальная и средняя относительная погрешность восстановления общего содержания $CO_2$ из модельных данных методами ЭОФ, НС и СД, %**

| Метод | МОП* | СОП* |
|---|---|---|
| ЭОФ | 0,097 | 0,032 |
| НС | 0,077 | 0,018 |
| СД | 0,089 | 0,021 |

*МОП – Максимальная относительная погрешность, СОП – средняя относительная погрешность.

Сравнение результатов восстановления (см. таблицу) показывает, что все три непараметрических подхода практически одинаковы по своим характеристикам (скорость, точность, устойчивость). Несколько лучшие характеристики, по точности, показывает метод нейронных сетей.

**Заключение**

В статье приводится описание непараметрических подходов (метод эмпирических ортогональных функций, нейронных сетей и случайных деревьев) и результаты восстановления общего содержания $CO_2$ из модельных данных измерений спутниковым прибором GOSAT. Численное моделирование решения обратной задачи показывает, что подходы сопоставимы между собой по точности восстановления и скорости обработки данных измерений. Дальнейшее развитие подходов связано с обработкой реальных спутниковых данных.

*Литература*

1. Малкевич М.С. Оптические исследования атмосферы со спутников. – М.: Наука, 1972. – 303 с.
2. Global Concentrations of CO2 and CH4 Retrieved from GOSAT: First Preliminary Results / T. Yokota, Y. Yoshida, N. Eguchi et all. // SOLA. – 2009. – Vol. 5. – P. 160–163.
3. Ляхов А.Н. Современные методы обработки данных в геофизике // Иркутск: Лекции БШФФ, 2006. – С. 39–46.
4. Хайкин С. Нейронные сети: полный курс. – М.: Вильямс, 2006. – 1104 с.
5. Осовский С. Нейронные сети для обработки информации. – М.: Финансы и статистика, 2002. – 344 с.
6. Kataev M.Yu. Comparison of the nonparametric approaches to retrieving of CO2 total content from satellites measurements / M.Yu. Kataev, S.G. Kataev, A.G. Andreev, S.A. Bazelyuk // International Conference Computational Information Technologies for Environmental Sciences (CITES-2011), 9–11 July 2011, Tomsk, Russia. – Tomsk: 2011.
7. Breiman L. Random Forests // Mach. Learn. – 2001. – Vol. 45, № 1. – P. 5–32.
8. Pal M. Random forest classifier for remote sensing classification // International Journal of Remote Sensing. – 2005. – Vol. 26. – P. 217–222.
9. Kataev M.Yu. Information-processing software for satellite signal modeling in global scale / M.Yu. Kataev, A.K. Lukianov // International Conference on Environmental Observations, Modeling and Information Systems (ENVIROMIS-2010), 5–11 July 2010, Tomsk, Russia. – Tomsk: 2010.

---

**Катаев Михаил Юрьевич**
Д-р техн. наук, профессор каф. автоматизированных систем управления (АСУ) ТУСУРа
Тел.: 8 (382-2) 70-15-36
Эл. почта: kataev.m@sibmail.com

**Катаев Сергей Григорьевич**
Канд. физ.-мат. наук, доцент, докторант каф. АСУ ТУСУРа
Тел.: 8 (382-2) 70-15-36

**Андреев Алексей Геннадьевич**
Магистрант каф. АСУ ТУСУРа
Тел.: 8 (382-2) 70-15-36

**Базелюк Сергей Андреевич**
Магистрант каф. АСУ ТУСУРа
Тел.: 8 (382-2) 70-15-36

**Лукьянов Андрей Кириллович**
Аспирант каф. АСУ ТУСУРа
Тел.: 8 (382-2) 70-15-36

Kataev M.Yu., Kataev S.G., Andreev A.G., Bazeluyuk S.V., Lukyanov A.K.
**Nonparametric approaches retrieving of the $CO_2$ total content from data of satellite monitoring**

In article is given the comparing of nonparametric approaches for solving problem of retrieving of $CO_2$ total content on measurement data by the GOSAT satellite device. The results of test calculations of reflected from surface solar radiation in near IR spectral region are presented.
**Keywords:** reflected from surface solar radiation, satellite spectrometer, neural-network, random forests, empirica1 orthogonal function.