Динамическое формирование и обновление карты запасов органического углерода на территории России как задача интеллектуального анализа Больших данных

О. М. Голозубов[1], О. В. Чернова[2]

Рассматриваются принципы динамического расчета показателей и некоторые алгоритмы интеллектуального анализа данных (data mining), использованные при расчетах карт секвестрации и запасов органического углерода в почвах России в рамках проектов ФАО ООН по созданию глобальных карт. Приводится описание разномодальных и разновременных исходных данных: растровых сеток различных разрешений, векторных данных в географической системе координат, атрибутивной информации. Описан расчет итоговых карт и карт погрешностей в распределенной сети почвенных дата-центров как задачи BigData.

Ключевые слова: почвенные базы данных, статистические методы, распределенные системы, органический углерод.

В нашей стране за более чем десятилетнюю историю развития Информационной Системы «Почвенно-географическая база данных Российской Федерации» (ИС ПГБД РФ – https://soil-db.ru/) накоплен достаточный объем данных для решения фундаментальных и прикладных задач почвенного мониторинга. Большой массив почвенной информации накоплен научными учреждениями, региональными центрами агрохимической службы страны, а также другими организациями, которые осуществляют мониторинг и наполнение баз актуальных и архивных данных, выполняют обработку данных дистанционного зондирования (ДДЗ).

Ниже приводится краткая характеристика наборов почвенных данных, аккумулированных в ИС ПГБД:

[1] *Голозубов Олег Модестович* — ведущий научный сотрудник, к.б.н., факультет почвоведения Московского государственного университета им. М.В. Ломоносова, e-mail: oleggolozubov@soil.msu.ru.

Golozubov Oleg Modestovich — leading researcher, PhD on biological sciences, Soil Science Faculty, Lomonosov Moscow State University

[2] *Чернова Ольга Владимировна* — старший научный сотрудник, к.б.н., Институт проблем экологии и эволюции им. А.Н. Северцова РАН, e-mail: ovcher@mail.ru.

Chernova Olga Vladimirovna — senior researcher, PhD on biological sciences, A.N. Severtsov Institute of Ecology and Evolution

Профильных данных: 10 500 профилей, из них на территории России – 3 000, в том числе представительных профилей с полным описанием – 900, и соответственно 22 000, 13 000 и 5 000 описаний горизонтов.

Мелкомасштабные почвенные карты и карты почвенно-экологического районирования (от М 1:1 000 000 и более мелкомасштабных): 36 000 контуров, из них на территории России 27 000;

Мелкомасштабные тематические карты: климатические, растительности, лесов, почвообразующих пород, природно-сельскохозяйственного районирования и другие.

Среднемасштабные почвенные карты, почвообразующих пород, эрозии (от М 1:200 000 до М 1:600 000): 30 000 контуров, из них на территории России 27 000.

Крупномасштабные почвенные карты: 287 000 контуров, из них на территории России 184 000 контуров, более 25 000 000 га.

Также ИС ПГБД аккумулирует большой объем сопутствующей информации: данные агрохимических обследований, геоботанические и геоморфологические описания, карты землеустройства, севооборотов, эрозии и негативных факторов, справочники методов измерения, классификаторов и многое другое.

Архитектурно-организационные принципы для построения информационных систем почвенного мониторинга и, более широко, пространствен-но-природных данных сочетают в себе как типовые принципы Big Data, так и особенности пространственно-распределенных мониторинговых сетей. Так, организационным комитетом INSPIRE («Инфраструктура пространственной информации в Евросоюзе»), образованным в 2007 г., проведена работа для создания информационной инфраструктуры, обеспечивающей свободный публичный доступ к пространственной природно-экологической информации [2]. В ней предусматривается: а) хранение только первичных данных в одном месте сбора и обработки; б) обеспечение неразрывности пространственных данных на административных границах; в) их доступности в различных масштабах (от детального для исследований, до обзорного для стратегических целей), и г) обеспечивается «прозрачность» поиска нужной пространственной информации при понятных условиях доступа к ней.

Задачи почвенного мониторинга можно отнести к классу задач Big Data [1]. Основной объект приложения Big Data – искусственные среды (экономика, торговля, курсы валют), или «организованные» среды (здравоохранение), в которых решены проблемы получения исходных данных в цифровой форме. Для анализа естественных природных сред – в метеорологии, геологии, экологии и почвоведении – требуется и более полный сбор данных, и их «гармонизация», и включение моделей в интерпретацию и прогноз.

Проект INSPIRE, так же как ИС ПГБД РФ и аналогичные глобальные мониторинговые проекты решают проблемы постоянного привлечения больших объемов новой информации, ее хранения и обработки. Система управления базами данных (СУБД) и приложения баз данных организуются таким образом, чтобы минимизировать пересылки данных по сети, связывающей узлы соответствующей вычислительной системы. Достигается реальное распараллеливание работы СУБД и приложений, поскольку при отсутствии общих ресурсов между узлами вычислительной системы уменьшается вероятность конфликтов между частями системы и приложений, выполняемыми в разных узлах сети. Также обеспечивается поддержка оперативной аналитической обработки данных [3]. На рис. 1 приведена организационная схема распределенной базы данных ИС ПГБД РФ, отражающая указанные выше принципы.

Рис. 1. Организационная схема ИС ПГБД РФ

Типовые методы и механизмы систем Big Data применительно к задачам почвенного мониторинга имеют ряд особенностей. Исходно, технология комплексного оперативного многомерного анализа данных получила название OLAP. OLAP – это ключевой компонент организации хранилищ данных. Концепция OLAP была описана в 1993 году Эдгаром Коддом, известным исследователем баз данных и автором реляционной модели данных [4]. В литературе эти технологии часто называют Data mining - методы обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Для принципиально неполных и разреженных почвенных

данных прогнозирование – важный этап экологического мониторинга, позволяющий заполнить пробелы в данных. При этом решаются задачи регрессии и классификации, формируются сводные отчеты на многомерном «кубе» и глобальные многослойные географические карты, в основном здесь применяются «облачные» решения, поскольку исходная информация тематически и пространственно распределена между соответствующими узлами сети.

В 2017 году ФАО ООН инициировал проект по созданию всемирной карты запасов органического углерода в 30-сантиметровом слое почвы (GSOC17) [5], была предложена единая методика расчетов, при реализации которой требовалось составить также карту оценки погрешности расчетов. В 2020 году в связи с проблемами изменения климата ФАО предложило сформировать мировую карту секвестрации органического углерода SOCSeq, основанную на обновленной карте GSOC17. Обе карты представляют собой постоянно обновляемые веб-ресурсы (https://www.fao.org/global-soil-partnership/pillars-action/4-information-and-data-new/global-soil-organic-carbon-gsoc-map/en/) – рис.2.

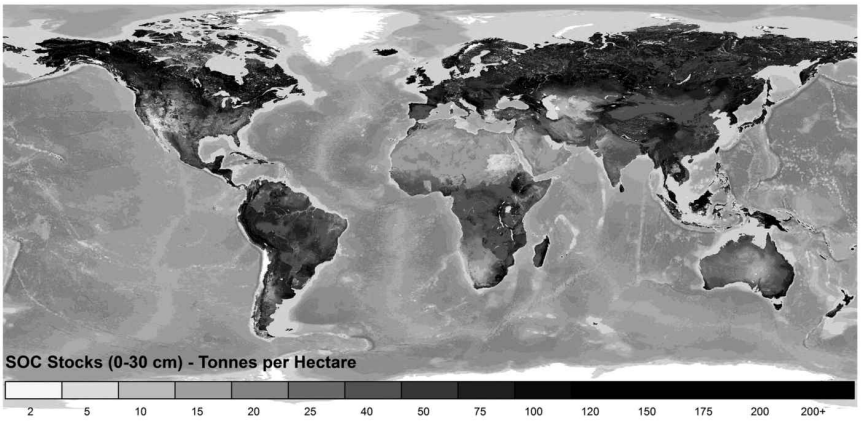

Рис. 2. Пример карты запасов органического углерода на сайте ФАО ООН

Для расчета карт, наряду с прочими характеристиками, были необходимы сведения о редко определяемой экспериментально объемной массе почвенных горизонтов (bulk density). Кроме того, в методике одним из параметров расчета является процентное содержание частиц <0.002 мм в слое почвы 0-30 см [6], что отличается от данных о гранулометрическом составе в градациях, принятых в отечественном почвоведении. Для расчета объемной массы минеральных горизонтов использовали предложенную О.Г. Честных и Д.Г. Замолодчиковым пятипараметрическую функцию нелинейной регрессии, которая позволяет прогнозировать объ-

емную массу в зависимости от содержания гумуса и глубины горизонта [7].

Применимость уравнения регрессии была проверена на заведомо независимых информационных массивах. Верификация уравнения и коэффициентов для почв группы «Таежные» проведена на основе данных по 125 горизонтам из 31 разреза дерново-подзолистых почв Московской области, преимущественно глееватых и глеевых. Средняя относительная ошибка определения плотности для этих почв составила 7.5% (рис. 3). Для почв группы «Степные» применимость коэффициентов проверяли на характеристиках 307 горизонтов из 111 разрезов черноземов обыкновенных и южных черноземов Ростовской области (относительная ошибка - 7.6%) [8].

Для расчета процентного содержания частиц гранулометрической фракции <0.002 была применена модель, состоящая из последовательного ряда статистических методов, в том числе регрессионного анализа, в рамках которой:

- данные представительных профилей, содержащие минимально необходимый набор показателей (гранулометрический состав и содержание гумуса), усреднялись по принадлежности к одинаковым группам почв;
- для приведения полученных по горизонтам данных к слою 0-30 см рассчитывали средневзвешенные значения показателей.

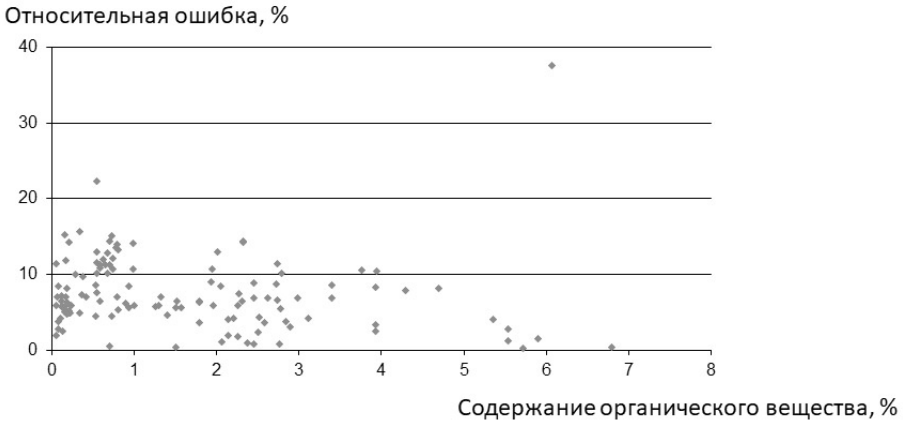

Рис. 3. Зависимость относительных ошибок расчета объемной массы дерново-подзолистых почв от их обогащенности органическим веществом

Такие общепринятые механизмы Big Data как Map Reduce также находят своё применение в мониторинговых системах. Задача распаралле-

ливания данных и решений (Map) здесь решается естественным путем, поскольку вычисления над природно-почвенными данными как правило затрагивают географически локальную информацию и тематически гетерогенную, и размещенную в соответствующих распределенных узлах. А задача сокращения вычислений и их типизация (Reduce) выполняется за счет а) двухтактной системы сбора данных – сначала метаданные, а затем собственно данные от определенных узлов, и б) предобработки данных в узлах сети и окончательной сборки в «облаке».

Список литературы

[1] Leskovec J., Rajaraman A., Ullman J., *Mining of Massive Datasets (3rd ed.)*, Cambridge University Press, Cambridge, 2020.

[2] "INSPIRE D2.8.III.3 Data Specification on Soil – Technical Guidelines", https://inspire.ec.europa.eu/data-specifications/2892.

[3] Golozubov O.M., Rozhkov V.A., Alyabina I.O., Ivanov A.V., Kolesnikova V.M., Shoba S.A., "Technologies and Standards in the Information Systems of the Soil-Geographic Database of Russia", *Eurasian Soil Science*, **48**:1 (2015), 1–10.

[4] Codd E.F., Codd S.B., Salley C.T., "Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate. Technical report", 1993.

[5] FAO, *A protocol for measurement, monitoring, reporting and verification of soil organic carbon in agricultural landscapes – GSOC-MRV Protocol*, FAO, Rome, 2020.

[6] FAO, *Technical specifications and country guidelines for Global Soil Organic Carbon Sequestration Potential Map (GSOCseq)*, FAO, Rome, 2020.

[7] Chestnykh O.V. and Zamolodchikov D.G., "Bulk density of soil horizons as dependent on their humus content", *Eurasian Soil Science*, **37** (2004), 816–823.

[8] Chernova O.V., Golozubov O.M., Alyabina I.O. et al., "Integrated Approach to Spatial Assessment of Soil Organic Carbon in the Russian Federation", *Eurasian Soil Science*, **54** (2021), 325–336.

Dynamic formation and updating of the map of organic carbon stock in Russia as a task of Big Data mining

Golozubov O.M., Chernova O.V.

The principles of dynamic calculation of indicators and some algorithms of data mining used in the calculation of maps of sequestration and stock of organic carbon in the soils of Russia in the framework of FAO UN projects to create global maps are considered. The description of multi-modal and multi-temporal source data is given: raster grids of various resolutions, vector data in the geographical coordinate system, attribute information. The calculation of final maps and error maps in a distributed network of soil data centers is described as a Big Data task.

Keywords: soil databases, statistical methods, distributed systems, organic carbon.

References

[1] Leskovec J., Rajaraman A., Ullman J., *Mining of Massive Datasets (3rd ed.)*, Cambridge University Press, Cambridge, 2020.

[2] "INSPIRE D2.8.III.3 Data Specification on Soil – Technical Guidelines", https://inspire.ec.europa.eu/data-specifications/2892.

[3] Golozubov O.M., Rozhkov V.A., Alyabina I.O., Ivanov A.V., Kolesnikova V.M., Shoba S.A., "Technologies and Standards in the Information Systems of the Soil-Geographic Database of Russia", *Eurasian Soil Science*, **48**:1 (2015), 1–10.

[4] Codd E.F., Codd S.B., Salley C.T., "Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate. Technical report", 1993.

[5] FAO, *A protocol for measurement, monitoring, reporting and verification of soil organic carbon in agricultural landscapes – GSOC-MRV Protocol*, FAO, Rome, 2020.

[6] FAO, *Technical specifications and country guidelines for Global Soil Organic Carbon Sequestration Potential Map (GSOCseq)*, FAO, Rome, 2020.

[7] Chestnykh O.V. and Zamolodchikov D.G., "Bulk density of soil horizons as dependent on their humus content", *Eurasian Soil Science*, **37** (2004), 816–823.

[8] Chernova O.V., Golozubov O.M., Alyabina I.O. et al., "Integrated Approach to Spatial Assessment of Soil Organic Carbon in the Russian Federation", *Eurasian Soil Science*, **54** (2021), 325–336.