Геофизические технологии. 2024. № 1. С. 6–18. **www.rjgt.ru**

Russian Journal of Geophysical Technologies. 2024. No. 1. P. 6–18.

Научная статья / Original article

УДК 550.34.013.4

doi:10.18303/2619-1563-2024-1-6

АДАПТИВНАЯ ОПТИМИЗАЦИЯ ОБУЧАЮЩЕЙ ВЫБОРКИ ПРИ НЕЙРОСЕТЕВОМ ПОДХОДЕ К ПОДАВЛЕНИЮ ЧИСЛЕННОЙ ДИСПЕРСИИ В ДАННЫХ СЕЙСМИЧЕСКОГО МОДЕЛИРОВАНИЯ

К.А. Гадыльшина, В.В. Лисица✉, К.Г. Гадыльшин, Д.М. Вишневский, В.И. Костин

Институт нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН,
630090, Новосибирск, просп. Акад. Коптюга, 3, Россия,
✉*Вадим Викторович Лисица, LisitsaVV@ipgg.sbras.ru, https://orcid.org/0000-0003-3544-4878*

Аннотация. Представлен новый подход к построению обучающей выборки для NDM-net (Numerical dispersion mitigation neural network) – искусственной нейронной сети, применяющейся для подавления численной ошибки в результатах численного сейсмического моделирования. На первом этапе небольшое количество сейсмограмм, рассчитанных с использованием грубой и мелкой сеток, используется для обучения сети, сопоставляющей неточные данные, полученные в результате расчета на крупной сетке, с высококачественными данными с мелкой сетки. Затем сеть NDM-net обрабатывает весь набор данных, предварительно рассчитанных с использованием грубой сетки, для уменьшения численной ошибки. Самая трудоемкая часть предлагаемого алгоритма – генерация набора обучающих данных. Возникает необходимость минимизировать количество сейсмограмм в наборе обучающих данных без потери качества обучения. Выбор обучающих данных осуществляется с фиксацией расстояния Хаусдорфа между набором обучающих данных и всем набором данных. При этом уровень предельного расстояния варьируется в зависимости от используемой для моделирования сейсмогеологической модели. Показано, что адаптивная стратегия предпочтительнее фиксированного ограничения метрики Хаусдорфа, поскольку она позволяет сократить набор обучающих данных без потери точности работы обученной сети NDM-net.

Ключевые слова: сейсмическое моделирование, численная дисперсия, глубокое обучение

Финансирование: работа выполнена при поддержке Российского научного фонда, грант № 22-11-00004.

Для цитирования: Гадыльшина К.А., Лисица В.В., Гадыльшин К.Г., Вишневский Д.М., Костин В.И. Адаптивная оптимизация обучающей выборки при нейросетевом подходе к подавлению численной дисперсии в данных сейсмического моделирования // Геофизические технологии. 2024. № 1. С. 6–18. doi:10.18303/2619-1563-2024-1-6.

ADAPTIVE TRAINING DATASET GENERATION FOR NEURAL NETWORK NUMERICAL DISPERSION MITIGATION APPROACH IN SEISMIC MODELING

K.A. Gadylshina, V.V. Lisitsa✉, K.G. Gadylshin, D.M. Vishnevsky, V.I. Kostin

Trofimuk Institute of Petroleum Geology and Geophysics SB RAS, Koptyug Ave., 3, Novosibirsk, 630090, Russia,
✉*Vadim V. Lisitsa, LisitsaVV@ipgg.sbras.ru, https://orcid.org/0000-0003-3544-4878*

Abstract. We introduce a novel method for developing the training dataset for the Numerical Dispersion Mitigation network (NDM-net), aimed at diminishing numerical inaccuracies in seismic modeling. Our strategy involves using a limited set of seismograms, produced with coarse and fine grids, to train the network. This training enables the network to transform less accurate coarse-grid data into higher-quality fine-grid data. Subsequently, the network is employed on a more extensive set of seismograms, initially computed with the coarse grid, to lower numerical errors. Creating the training dataset is the

most demanding aspect of this method, requiring a balance between the number of seismograms used and maintaining training effectiveness. We propose a method to create the training dataset that maintains a specific Hausdorff distance with the complete dataset. However, this distance can vary based on the seismic-geological model used in simulations. Our work shows that an adaptive approach in setting the Hausdorff distance limit is more advantageous than a fixed limit, as it reduces the training dataset size without compromising accuracy.

Keywords: seismic modeling, numerical dispersion, deep learning

Funding: the work was supported by the Russian Science Foundation, Project No. 22-11-00004.

For citation: Gadylshina K.A., Lisitsa V.V., Gadylshin K.G., Vishnevsky D.M., Kostin V.I. Adaptive training dataset generation for neural network numerical dispersion mitigation approach in seismic modeling // Russian Journal of Geophysical Technologies. 2024. No. 1. P. 6–18. (In Russ.). doi:10.18303/2619-1563-2024-1-6.

ВВЕДЕНИЕ

Сейсмическое моделирование требует решения волнового уравнения или системы уравнений динамической теории упругости [Virieux et al., 2011]. Известно, что численная ошибка получаемого решения напрямую зависит от шага расчетной сетки или, что тоже самое, от размера решаемой задачи. Так при использовании конечно-разностных схем на сдвинутых сетках четного порядка аппроксимации [Virieux, 1986; Levander, 1988] численная ошибка проявляется в виде дисперсии сигнала. Как правило, сетка строится таким образом, чтобы временная задержка из-за дисперсии записываемого сигнала не превышала четверти периода волны. Это условие должно удовлетворяться для сигнала, распространяющегося на 100–200 длин волн, таким образом для типичных систем наблюдения пространственный шаг сетки составляет около 1 м. С учетом размера стандартной области моделирования 10 × 10 × 5 километров требуется приблизительно 5×10^{11} точек сетки для дискретизации модели или 5×10^{12} степеней свободы для численного моделирования волнового поля в изотропных упругих средах. Действительно, современные суперкомпьютеры способны выполнять такие вычисления. В то же время эта оценка справедлива для моделирования волнового поля, соответствующего только одному источнику, тогда как типичная система наблюдения содержит около 10^5 сейсмических источников. Таким образом, моделирование всего набора данных требует неприемлемо большого объема вычислительных ресурсов.

Один из подходов к уменьшению численной ошибки основан на постобработке смоделированных данных. В том числе прямое вычитание дисперсии [Koene et al., 2017; Mittet, 2019]. Этот подход позволяет частично подавлять дисперсию, вызванную аппроксимацией производных по времени, а дисперсия, обусловленная аппроксимацией эллиптической части оператора, не уменьшается. Другое направление исследований в этой области – применение машинного обучения [Siahkoohi et al., 2019; Gadylshin et al., 2022b]. В частности, для обработки сейсмограмм, построенных по рассчитанным при численном моделировании на грубой сетке параметрам, применяется уменьшающая численную дисперсию искусственная нейронная сеть NDM-net (Numerical dispersion mitigation neural network – в англоязычной литературе) [Gadylshin et al., 2022b]. Подход с применением NDM-net основан на необходимости решения множества аналогичных задач с разными правыми частями (для разных положений источников сигнала). В результате набор обучающих данных строится как точное решение для небольшого числа задач, а затем сеть применяется для обработки всего набора данных. Однако формирование набора обучающих данных

наиболее трудозатратно, т. к. предполагает моделирование упругого волнового поля с использованием достаточно мелкой сетки. Главным образом, чтобы улучшить производительность алгоритма, необходимо минимизировать набор обучающих данных. В соответствии с этим требованием, набор обучающих данных строится с фиксацией конкретного значения метрики Хаусдорфа между набором обучающих данных и всем набором данных [Gadylshin et al., 2022a]. Итак, количество сейсмограмм в обучающем наборе данных сокращается в три раза по сравнению с вариантом отбора сейсмограмм от равномерно распределенных источников. Тем не менее, если модель имеет сильные латеральные неоднородности, такие как дайки или соляные интрузии, среднее расстояние между сейсмограммами варьируется. В связи с этим для некоторых участков области моделирования возможен выбор разреженного набора источников без снижения представительности набора, что влечет необходимость локального увеличения порогового значения метрики Хаусдорфа. Вместе с тем, из-за изменения порогового значения в большую сторону для всего набора, данные из относительно гладких областей модели утрачивают репрезентативность. В связи с вышеизложенным предлагается адаптивный выбор набора обучающих данных, при котором предельное расстояние между набором обучающих данных и всем набором данных изменяется в соответствии с моделью.

Предварительные замечания

Известно, что в настоящее время наиболее распространенный метод для выполнения сейсмического моделирования – конечно-разностный метод, сочетающий в себе высокую вычислительную эффективность с простотой реализации [Lisitsa et al., 2010; Virieux et al., 2011; Vishnevsky et al., 2014]. Для моделирования волновых полей в изотропных упругих средах в двумерном случае используются схемы на сдвинутых сетках [Virieux, 1986; Levander, 1988]:

$$
\begin{gathered}
\hat{\rho}_{i+\frac{1}{2},j} D_t[u_x]^{n+\frac{1}{2}}_{i+\frac{1}{2},j} = D_x[\sigma_{xx}]^{n+\frac{1}{2}}_{i+\frac{1}{2},j} + D_z[\sigma_{xz}]^{n+\frac{1}{2}}_{i+\frac{1}{2},j}, \\
\hat{\rho}_{i+\frac{1}{2},j} D_t[u_z]^{n+\frac{1}{2}}_{i,j+\frac{1}{2}} = D_x[\sigma_{xz}]^{n+\frac{1}{2}}_{i,j+\frac{1}{2}} + D_z[\sigma_{zz}]^{n+\frac{1}{2}}_{i,j+\frac{1}{2}}, \\
D_t[\sigma_{xx}]^n_{i,j} = (\lambda + 2\mu)_{i,j} D_x[u_x]^n_{i,j} + \lambda_{i,j} D_z[u_z]^n_{i,j} + (f_{xx})^n_{i,j}, \\
D_t[\sigma_{zz}]^n_{i,j} = \lambda_{i,j} D_x[u_x]^n_{i,j} + (\lambda + 2\mu)_{i,j} D_z[u_z]^n_{i,j} + (f_{zz})^n_{i,j}, \\
D_t[\sigma_{xz}]^n_{i+\frac{1}{2}\,j+\frac{1}{2}} = \hat{\mu}_{i+\frac{1}{2},j+\frac{1}{2}} \left(D_x[u_z]^n_{i+\frac{1}{2},j+\frac{1}{2}} + D_z[u_x]^n_{i+\frac{1}{2},j+\frac{1}{2}} \right) + (f_{xz})^n_{i+\frac{1}{2},j+\frac{1}{2}},
\end{gathered} \tag{1}
$$

где $\boldsymbol{u} = \left(u_x, u_y\right)^T$ – вектор скорости, σ_{xx}, σ_{zz} и σ_{xz} – компоненты тензора напряжений. Соответствующие сеточные функции определяются на сдвинутой сетке, так что $g^N_{I,J} = g(N\tau, Ih_x, Jh_z)$, где τ – шаг по времени, h_x и h_z – шаги по пространству, индексы I, J, N могут быть как целыми, так и полуцелыми, а g – достаточно гладкая функция. Функции f_{xx}, f_{zz} и f_{xz} представляют правую часть. Параметр ρ – плотность среды, λ и μ – параметры Ламе. Как правило, все параметры модели определяются в точках целочисленной сетки. При этом для их вычисления в дробных точках используются среднее арифметическое значение для плотности и среднее гармоническое значение для μ [Moczo et al., 2002; Vishnevsky et al., 2014]. Операторы D_t, D_x и D_z:

$$D_t[g]_{I,J}^N = \frac{g_{I,J}^{N+\frac{1}{2}} - g_{I,J}^{N-\frac{1}{2}}}{\tau} = \frac{\partial g}{\partial t} + O(\tau^2),$$

$$D_x[g]_{I,J}^N = \frac{1}{h_x}\sum_{m=0}^{M}\left(g_{I+m+\frac{1}{2},J}^N - g_{I-m-\frac{1}{2},J}^N\right) = \frac{\partial g}{\partial x} + O\left(h_x^{2(m+1)}\right),$$

$$D_z[g]_{I,J}^N = \frac{1}{h_z}\sum_{m=0}^{M}\left(g_{I+m+\frac{1}{2},J}^N - g_{I-m-\frac{1}{2},J}^N\right) = \frac{\partial g}{\partial z} + O\left(h_z^{2(m+1)}\right). \quad (2)$$

Схема (1) аппроксимирует линейную гиперболическую систему, в связи с этим справедлив критерий устойчивости Куранта $\tau = Ch$, где $h = \min\{h_x, h_z\}$, и рассматривается схема, зависящая только от параметра h.

В кратких обозначениях задача в конечно-разностной постановке:

$$R_h[\boldsymbol{v}_h] = \phi_h(t)\delta(x - x_s^l)(z - z_s^l),$$

где $\phi(t)$ – временной импульс источника, δ – дельта-функция, (x_s^l, z_s^l) – координаты источника l. Вектор $\boldsymbol{v}_h$ – вектор решения, включающий компоненты вектора скорости и компоненты тензора напряжений.

Согласно теории конечно-разностных схем, если схема устойчива и аппроксимирует исходный дифференциальный оператор, то численное решение сходится к истинному решению и имеет место следующая оценка:

$$\|\boldsymbol{v} - \boldsymbol{v}_h\| \le Ch^r, \quad (3)$$

где скорость сходимости r совпадает с порядком аппроксимации, который для рассматриваемого случая равен 2. Константа C не зависит от шага сетки. Тогда схема (1) с двумя разными шагами сетки $h_1 \le h_2$, следовательно согласно оценке (3):

$$\left\|\boldsymbol{v} - \boldsymbol{v}_{h_1}\right\| \le \left\|\boldsymbol{v} - \boldsymbol{v}_{h_2}\right\|.$$

Однако с уменьшением шага сетки $h_1 \le h_2$ размерность задачи увеличивается. Обучение сети NDM-net заключается в построении отображения

$$\mathcal{M}\left[\boldsymbol{v}_{h_2}\right] = \widetilde{\boldsymbol{v}}_{h_2},$$

такого, что

$$\left\|\widetilde{\boldsymbol{v}}_{h_2} - \boldsymbol{v}_{h_1}\right\| \le \varepsilon_{21} \ll Ch_2^r.$$

Если ошибка ε_{21} мала, тогда

$$\left\|\widetilde{\boldsymbol{v}}_{h_2} - \boldsymbol{v}\right\| \le \left\|\widetilde{\boldsymbol{v}}_{h_2} - \boldsymbol{v}_{h_1}\right\| + \left\|\boldsymbol{v}_{h_1} - \boldsymbol{v}\right\| \le \varepsilon_{21} + Ch_1^r < Ch_2.$$

Другими словами, NDM-net используется для сопоставления данных, вычисленных на грубой сетке, с данными, вычисленными с использованием достаточно мелкой сетки.

Как отмечалось выше, генерация обучающего набора данных наиболее трудоемка в реализации алгоритма с применением сети NDM-net, поскольку для некоторого количества источников волновое поле

моделируется с использованием мелкой сетки. Для уменьшения набора обучающих данных, т. е. для улучшения производительности алгоритма с применением сети NDM-net, необходимо максимально оптимизировать выбор репрезентативных источников из сети наблюдения.

ПОДХОДЫ К ПОСТРОЕНИЮ ОБУЧАЮЩЕЙ ВЫБОРКИ

Равномерно распределенные источники

Первый подход впервые предложен совместно с сетью NDM-net [Gadylshin et al., 2022b]. Пусть $\boldsymbol{u}^l$ – сейсмограмма для источника под номером l такая, что

$$\boldsymbol{u}^l = \boldsymbol{u}(t, x, z_0, x_s^l, z_s^l),$$

где $\boldsymbol{u}$ – вектор скорости. Тогда удобно сделать замену переменных $x_0 = x - x_s^l$, переменная x_0 характеризует смещение и фиксируется для всех исходных позиций. В результате сейсмические данные рассматриваются как набор двумерных функций от времени и смещения или как одна трехмерная функция, зависящая от времени, смещения и положения источника. Далее в выкладках используется первый вариант, таким образом, весь набор сейсмограмм как объединение

$$U = \cup_{l=1,\dots,L_s} \boldsymbol{u}(x_s^l, x_0, t) = \cup_{l=1,\dots,L_s} \boldsymbol{u}(\boldsymbol{x}_s^l),$$

и задача исследования – построить подмножество множества U:

$$U^{tr} = \bigcup_{l \in M_t} \boldsymbol{u}(\boldsymbol{x}_s^l) \subseteq U,$$

где $M_t \subseteq \{1, \dots, J_L\}$ – набор индексов сейсмограмм из подмножества U^{tr}.

Первый подход к построению набора данных U^{tr} реализуется для сохранения расстояния Хаусдорфа между набором обучающих данных и всем набором данных, где расстояние измеряется как физическое расстояние между позициями источников. Метрика Хаусдорфа между двумя множествами D_1 и D_2 определяется как

$$H(D_1, D_2) = \max\left\{\max_{i \in L_1} b(\boldsymbol{u}_i, D_2), \max_{j \in L_2} b(\boldsymbol{u}_j, D_1)\right\} = \max\left\{\max_{i \in L_1} \min_{j \in L_2} d(\boldsymbol{u}_i, \boldsymbol{u}_j), \max_{j \in L_2} \min_{i \in L_1} d(\boldsymbol{u}_j, \boldsymbol{u}_i)\right\},$$

где $b(\boldsymbol{u}_i, D_2) = \min_{j \in L_2} d(\boldsymbol{u}_i, \boldsymbol{u}_j)$ – расстояние между сейсмограммой $\boldsymbol{u}_i$ и множеством D_2, а $d(\boldsymbol{u}_i, \boldsymbol{u}_j)$ – расстояние между двумя элементами на множестве сейсмограмм. Тогда расстояние между набором обучающих данных $D_t \subset U$ и всем набором данных U определяется как

$$H(D_t, U) = \max_{i \in [1,\dots,J_s]} b(\boldsymbol{u}_i, D_t) = \max_{i \in [1,\dots,J_s]} \min_{j \in J_t} d(\boldsymbol{u}_i, \boldsymbol{u}_j).$$

Пусть $d(\boldsymbol{u}_i, \boldsymbol{u}_j) = |x_s^i - x_s^j|$ и при этом метрика Хаусдорфа полагается равной $M \cdot ds$, где $M \in \mathbb{Z}$, а ds – расстояние между двумя соседними источниками (обычно постоянное для всей системы сейсмической съемки). Таким образом весь набор обучающих данных U_s^N определен.

Расстояние между сейсмограммами на основе NRMS метрики

При втором подходе к построению набора обучающих данных метрика Хаусдорфа фиксируется, при этом вводится мера сходства между двумя сейсмограммами вместо опосредованной, измеряющей расстояние между источниками меры. В качестве меры подобия сейсмограмм используется нормализованное среднеквадратичное отклонение NRMS (Normalized Root Mean Square – в англоязычной литературе):

$$\mathrm{NRMS}(a_t, b_t, t_0) = \frac{200 \times \mathrm{RMS}(a_t - b_t)}{\mathrm{RMS}(a_t) + \mathrm{RMS}(b_t)},$$

где

$$\mathrm{RMS}(\phi_t) = \sqrt{\frac{\sum_{t_0-dt}^{t_0+dt} \phi_t^2}{N}},$$

где N – количество временных отсчетов в интервале $[t_0 - dt, t_0 + dt]$. Расстояние $d(\boldsymbol{u}^j, \boldsymbol{u}^k)$ здесь – среднее NRMS между сейсмограммами $\boldsymbol{u}^k$ и $\boldsymbol{u}^j$. Соответственно, определяется метрика Хаусдорфа и наборы обучающих данных таким образом, чтобы сохранить это расстояние. Такие наборы данных здесь обозначаются U^Q_{NRMS}, где Q – конкретное значение NRMS в процентах.

Для примера построения обучающего набора данных рассматривается соответствующая Восточной Сибири модель Ванавар (рис. 1). Прежде всего для этой модели рассчитываются параметры для двух наборов сейсмограмм на сетках с пространственными шагами 2.5 и 5.0 м соответственно. Все наборы данных содержат 1901 сейсмограмму с источниками, расположенными на расстоянии 100 м друг от друга. Затем отбираются обучающие данные и далее на основе NRMS вычисляются расстояния $b(\boldsymbol{u}_i, D_t)$ между каждой сейсмограммой и наборами обучающих данных в зависимости от положения источника (рис. 2). Для случая U^{70}_{NRMS} удается сохранить максимальное расстояние до набора обучающих данных, однако есть области, где используются почти все сейсмограммы, что существенно влияет на производительность алгоритма (рис. 3).

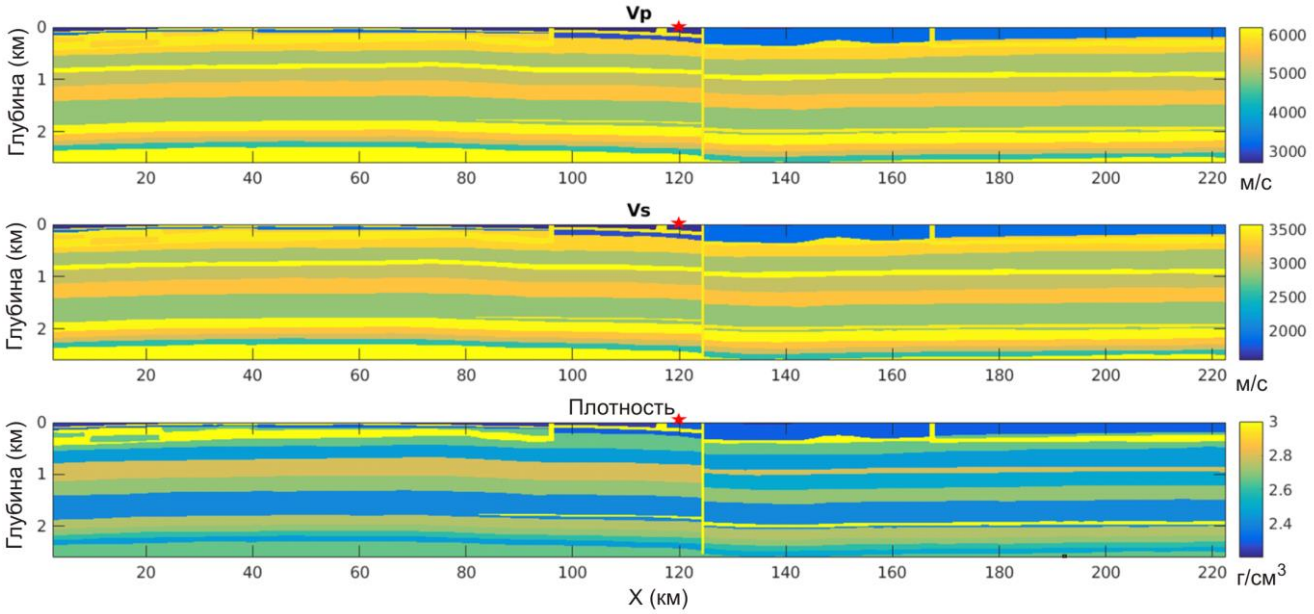

Рис. 1. Модель Ванавар: V_P – скорость продольной волны; V_s – скорость поперечной волны; плотность – ρ.

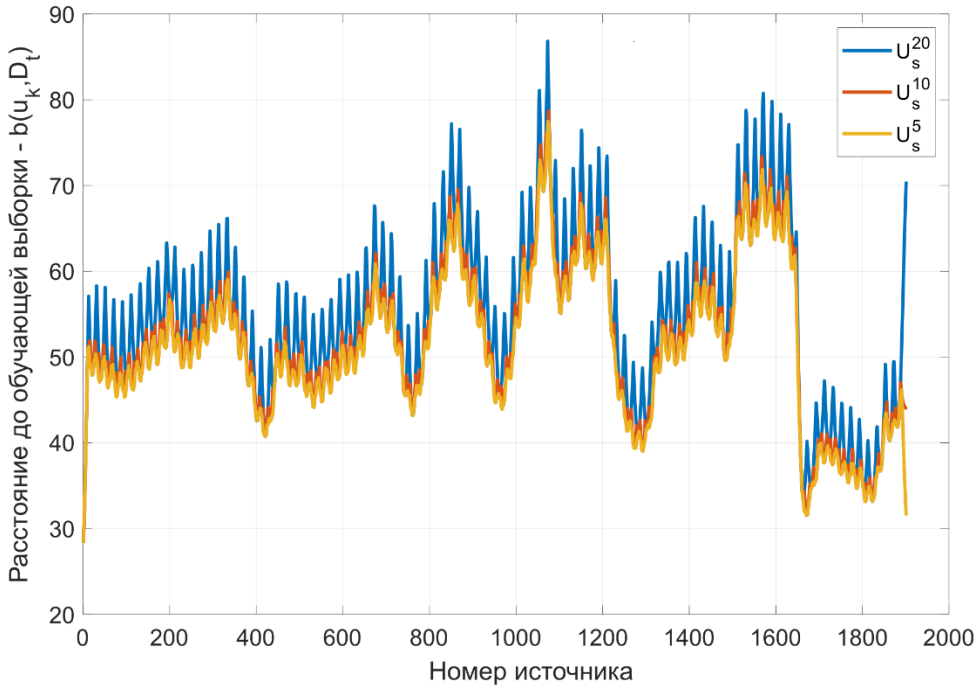

Рис. 2. Расстояния между сейсмограммами и обучающей выборкой $\mathrm{b}(\mathbf{u}_\mathrm{k}, \mathrm{U}_\mathrm{s}^\mathrm{N})$ для наборов равномерно распределенных источников U_s^5, $\mathrm{U}_\mathrm{s}^{10}$ and $\mathrm{U}_\mathrm{s}^{20}$.

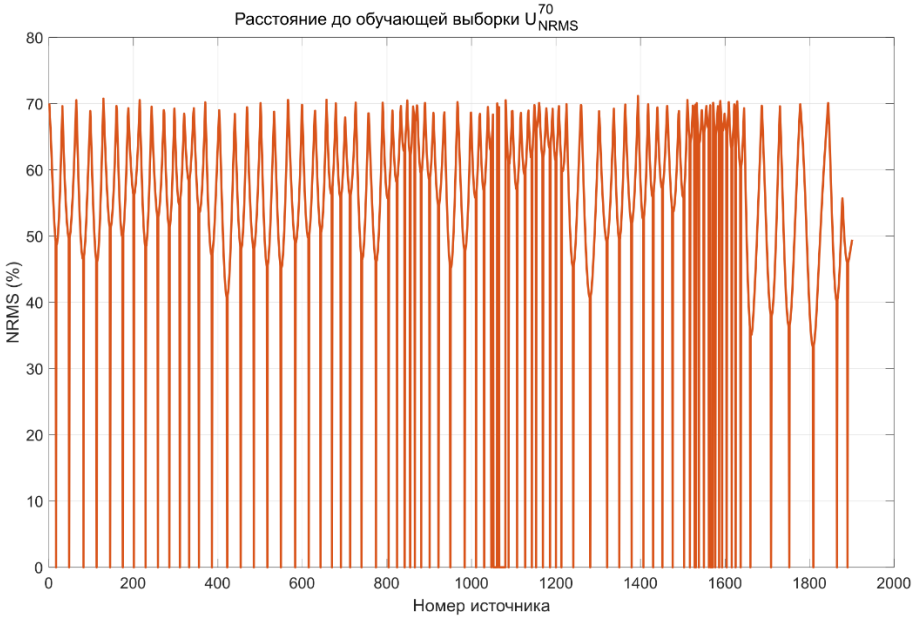

Рис. 3. Расстояние между сейсмограммами и обучающей выборкой $b(\boldsymbol{u}_k, {D_N}^{70}{}_{RMS})$.

Адаптивное построение набора данных

NRMS между сейсмограммами, соответствующими соседним источникам, изменяется от источника к источнику, и только формальное различие сейсмограмм не может быть надежным критерием для построения набора данных (рис. 2). Объединение двух вышеизложенных подходов позволяет выбрать предельное NRMS расстояние до набора обучающих данных в зависимости от физического положения источников. Вводится несколько уровней пределов метрики Хаусдорфа в терминах NRMS. Для этого рассчитывается среднее значение NRMS между сейсмограммами и его стандартное отклонение как

функция расстояния между источниками (рис. 4). Интервалы рассчитываются рекуррентно следующим образом:

- $L_1 =< \mathrm{NRMS}(1) > +\alpha\sigma(1)$, где $< \mathrm{NRMS}(1) >$ – среднее NRMS на расстоянии $1 \cdot ds$, и $\sigma(1)$ – стандартное отклонение от NRMS от источника на расстоянии $1 \cdot ds$, и α – изменяемый параметр;
- начало рекурсии;
- решение уравнения $\Delta_j = \mathrm{argmin}\left|L_{j-1} - < \mathrm{NRMS}(\Delta) >\right|$, где Δ_j– физическое расстояние между источниками;
- $L_j =< \mathrm{NRMS}(\Delta_j) > +\alpha\sigma(\Delta_j)$.

Процесс останавливается, если решением задачи argmin является максимум Δ. В алгоритме присутствует изменяемый параметр.

После определения предельных значений NRMS, обучающий набор данных отбирается по следующему алгоритму:

- Начало: сейсмограмма 1, подмножество 1;
- в порядке возрастания сейсмограммы набираются в текущий кластер, вместе с тем $\min_{i \in C_k} \max_{j \neq i} d(\boldsymbol{u}_i, \boldsymbol{u}_j) \leq L_m$, где i, j – индексы сейсмограмм текущего кластера C_k, а m – текущий предел NRMS (начинается с единицы для каждого нового кластера);
- если набор содержит достаточное количество сейсмограмм таких, что $\min_{i \in U_k} \max_{j \neq i} \left|x_s^i - x_s^j\right| \geq R$, то кластер определяется как полный, а следующий элемент обучающего набора данных – это решение уравнения $I_k = \mathrm{argmin}_{i \in C_k} \max_{j \neq i} d(\boldsymbol{u}_i, \boldsymbol{u}_j)$;
- если текущий набор содержит мало сейсмограмм таких, что $\min_{i \in U_k} \max_{j \neq i} \left|x_s^i - x_s^j\right| < R$, лимит NRMS полагается равным L_{m+1} и набор сейсмограмм в кластер продолжается.

Этот алгоритм разрешает ситуацию, когда в обучающем наборе данных оказываются сейсмограммы от слишком плотно расположенных источников.

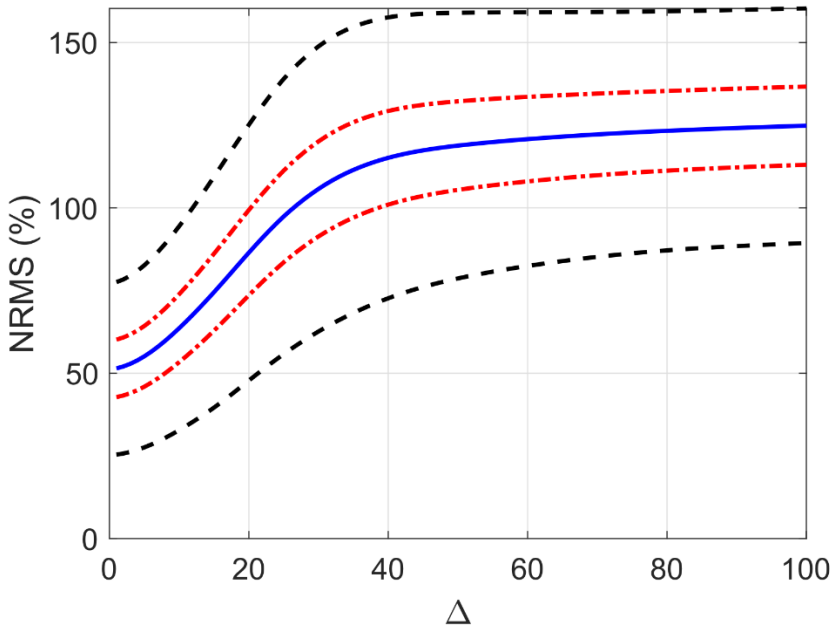

Рис. 4. Среднее значение (сплошная линия), среднеквадратичное отклонение (штрих-пунктирная линия) и тройное среднеквадратичное отклонение (пунктирная линия) NRMS-расстояние как функция расстояние между источниками.

ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ

Влияние параметров

Два параметра подбираются эмпирически: α – коэффициент, определяющий шаг роста предельного уровня NRMS, и R, определяющий минимальное количество сейсмограмм в кластере. Изменение минимального расстояния между источниками R напрямую влияет на количество сейсмограмм в наборе обучающих данных. В частности, чем выше R, тем меньше сейсмограмм будет использоваться. При этом имеет место следующая оценка $N_{cl} \leq N_{sg} \cdot ds/R$, где N_{cl} – количество кластеров или количество сейсмограмм в наборе обучающих данных, N_{sg} – общее количество сейсмограмм. Влияние α менее очевидно, поскольку α напрямую влияет на сегментацию предельного уровня NRMS, а не на количество сейсмограмм. Однако, чем меньше α, тем мельче дискретизация предельного NRMS. Таким образом, только физическая близость источников определяет набор обучающих данных. Рассматриваются следующие значения параметров $R \in \{2, 5, 10, 20, 50\}$ и $\alpha \in \{0.2, 0.5, 0.7, 0.9, 1.0, 1.2, 1.5, 2.0, 3.0\}$. Вычисляется зависимость между параметром α и количеством уровней сегментации предельной NRMS (табл. 1). Как и ожидалось, чем выше α, тем меньше уровней сегментации. В то же время, с увеличением α количество кластеров уменьшается (табл. 2). Именно из-за быстрого увеличения предельного уровня NRMS формируются более широкие кластеры сейсмограмм. Как указывалось выше, количество кластеров увеличивается с уменьшением минимального расстояния между источниками. В целом количество кластеров незначительно меняется при изменении предельного расстояния (по крайней мере, для нецелых значений α). Между тем при $R = 50 \cdot ds$ количество кластеров уменьшилось настолько, что не удалось выбрать репрезентативный набор обучающих данных.

Таблица 1

Число уровней сегментации для разных параметров α

α	0.2	0.5	0.7	0.9	1.0	1.2	1.5	2.0	3.0
N_L	30	12	9	7	6	5	5	4	3

Таблица 2

Количество сейсмограмм в обучающем наборе данных для различных параметров α и R

$R\backslash\alpha$	**0.2**	**0.5**	**0.7**	**0.9**	**1.0**	**1.2**	**1.5**	**2.0**	**3.0**
2	197	135	117	110	104	96	84	75	58
5	165	129	111	108	101	94	83	74	58
10	131	110	100	92	92	88	80	71	56
20	85	78	75	73	73	70	66	62	52
50	37	35	34	32	32	32	17	10	8

Реализация сети NDM-net

В последующих численных экспериментах $R = 10 \cdot ds$ в обоих случаях и $\alpha = 0.9$ и $\alpha = 1.2$. Сеть NDM-net обучается на двух адаптивных наборах данных, результаты сравниваются с результатами численных экспериментов, выполненных в предыдущих исследованиях, полученных с помощью обучающих наборов данных с равномерно распределенными источниками и с помощью метода сохранения NRMS. Гиперпараметры сети NDM-net одни и те же во всех численных экспериментах

[Gadylshin et al., 2022b]. Обученная сеть NDM-net применяется ко всему набору данных, результаты сравниваются с рассчитанным на мелкой сетке точным решением. Вычисляются расстояния NRMS по сейсмограммам между точными и прогнозируемыми решениями для наборов данных, построенных с сохранением NRMS и с использованием адаптивного метода соответственно (рис. 5, 6).

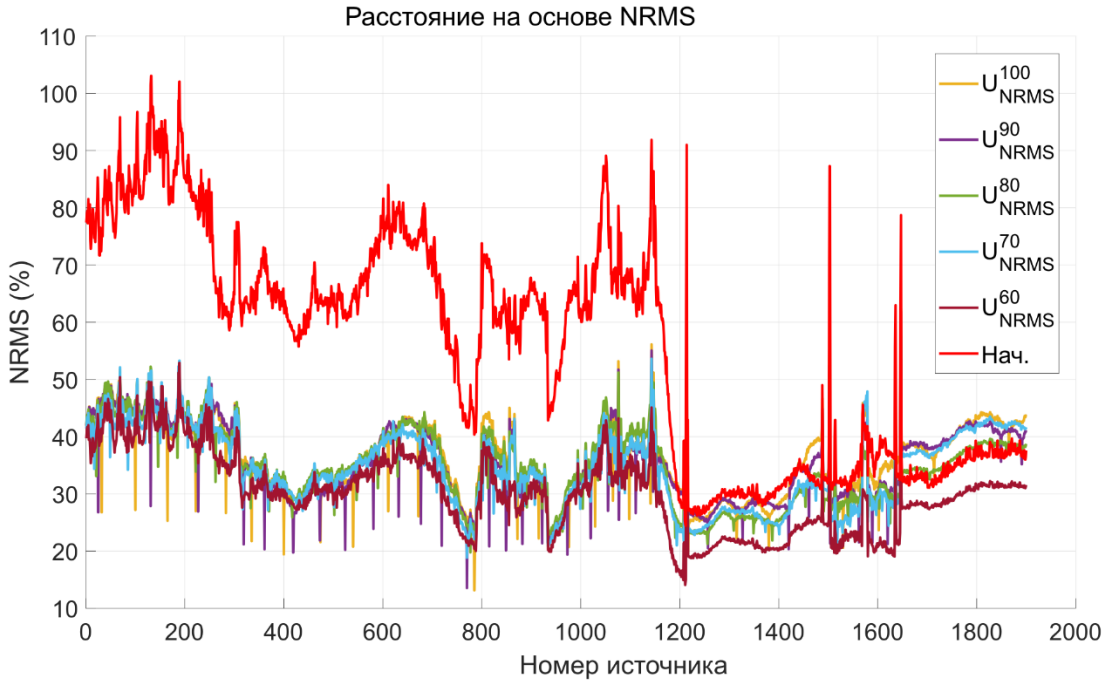

Рис. 5. Попарное NRMS расстояние между решением на мелкой сетке и результатом NDM-net постобработки для различных сценариев построения обучающих выборок на основе подхода с сохранением NRMS.

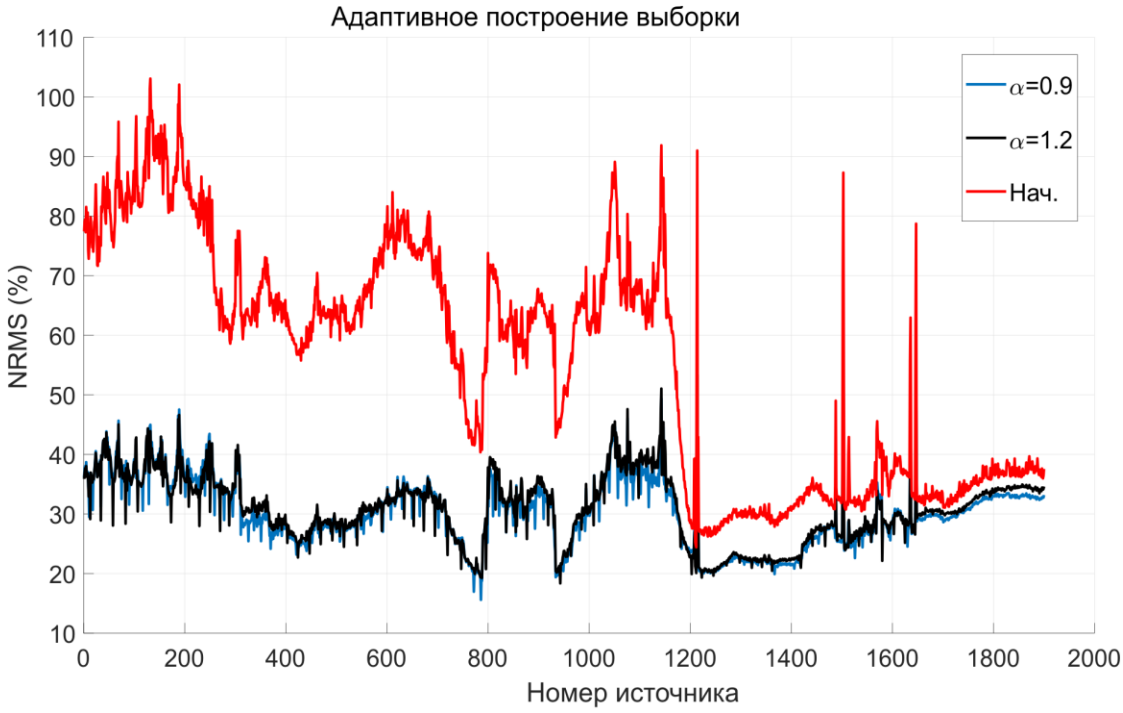

Рис. 6. Попарное NRMS расстояние между решением на мелкой сетке и результатом NDM-net постобработки для различных сценариев построения обучающих выборок на основе адаптивного подхода.

В двух рассматриваемых случаях уровень подавления численной дисперсии одинаковый. Для качественного анализа вычисляется среднее значение NRMS по всем сейсмограммам с учетом количества сейсмограмм в обучающих наборах данных. Кроме того, для полного сравнительного анализа учитываются результаты исследования с применением обучающего набора данных из сейсмограмм, соответствующих равномерно распределенным источникам (табл. 3, рис. 7). Результаты работы обученной на адаптивном наборе данных сети NDM-net превосходят остальные варианты и в высокой точности, так и в минимизации количества сейсмограмм, используемых для обучения.

Таблица 3

Различные сценарии выбора обучающего набора данных

Выбор обучающего набора данных с сохранением NRMS		
Набор данных	Количество сейсмограмм	Среднее NRMS
U_{NRMS}^{60}	414	28.16 %
U_{NRMS}^{70}	109	30.28 %
U_{NRMS}^{80}	56	34.69 %
U_{NRMS}^{90}	43	35.11 %
U_{NRMS}^{100}	34	35.68 %
Сейсмограммы от равномерно распределенных источников сигнала		
U_s^5	86	44.28 %
U_s^{10}	191	31.91 %
U_s^{20}	283	29.41 %
Адаптивный выбор обучающего набора данных		
$U^{\alpha=0.9}$	92	30.32 %
$U^{\alpha=1.2}$	88	30.38 %

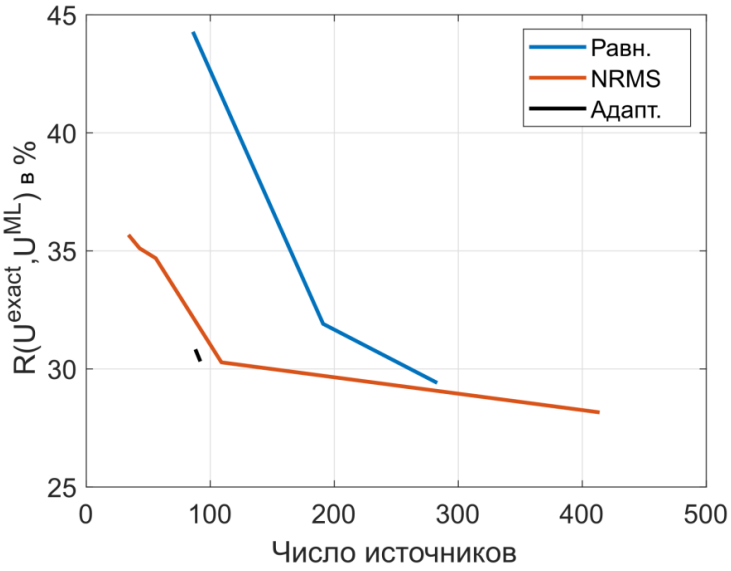

Рис. 7. Зависимость ошибки от номера сейсмограммы, использующаяся для различных стратегий построения обучающих выборок.

ВЫВОДЫ

Представлен новый способ построения набора обучающих данных для NDM-net. В отличие от подходов, где расстояние Хаусдорфа от набора обучающих данных до всего набора данных фиксировалось для всех сейсмограмм, новый подход адаптивен, что позволяет избежать использования неоправданно плотных наборов обучающих данных. В частности, в предыдущих исследованиях наборы данных были построены так, чтобы сохранить расстояние Хаусдорфа на основе евклидова расстояния между положениями источников. Ранее используемые обучающие наборы данных состояли из сейсмограмм, соответствующих эквидистантно распределенным источникам, без учета информации о степени схожести сейсмограмм. В дальнейшем расстояния в пространстве данных фиксируются. Для этого вводится NRMS мера сходства сейсмограмм. Так количество сейсмограмм в обучающем наборе данных сокращается в три раза по сравнению с эквидистантно распределенными источниками на модельном примере. Однако из-за локальных особенностей модели источники, соответствующие сейсмограммам из набора обучающих данных, распределены излишне плотно. Более того, для реализации отбора требуется эмпирический выбор входных параметров (предельного расстояния Хаусдорфа), которые значительно влияют на результат.

Предлагается корректировать расстояние до набора обучающих данных. В частности, рассматриваются сегментированные уровни ограничения NRMS на основе среднего значения NRMS для разных расстояний между источниками по всему набору данных. После этого применяется процедура построения обучающего набора данных с сохраненными NRMS. Если источники, соответствующие набору обучающих данных, близки друг к другу, то локально увеличивается уровень NRMS до предельного соответственно сегментированным уровням. Таким образом получаются наименьшие наборы обучающих данных среди трех рассмотренных методов их построения. Более того, при обучении на адаптивном обучающем наборе данных ошибка реализации сети NDM-net наименьшая из трех рассмотренных. Кроме того, при адаптивном подходе подбор входных параметров не требуется.

СПИСОК ИСТОЧНИКОВ / REFERENCES

Gadylshin K., Lisitsa V., Gadylshina K., Vishnevsky D. Optimization of the training dataset for numerical dispersion mitigation neural network // Lecture Notes in Computer Science. 2022a. Vol. 13378 LNCS. P. 295–309. doi:10.1007/978-3-031-10562-3_22.

Gadylshin K., Vishnevsky D., Gadylshina K., Lisitsa V. Numerical dispersion mitigation neural network for seismic modeling // Geophysics. 2022b. Vol. 87 (3). P. T237–T249. doi:10.1190/geo2021-0242.1.

Koene E.F.M., Robertsson J.O.A., Broggini F., Andersson F. Eliminating time dispersion from seismic wave modeling // Geophysical Journal International. 2017. Vol. 213 (1). P. 169–180. doi:10.1093/gji/ggx563.

Levander A.R. Fourth-order finite-difference *P-SV* seismograms // Geophysics. 1988. Vol. 53 (11). P. 1425–1436. doi:10.1190/1.1442422.

Lisitsa V., Podgornova O., Tcheverda V. On the interface error analysis for finite difference wave simulation // Computational Geosciences. 2010. Vol. 14 (4). P. 769–778. doi:10.1007/s10596-010-9187-1.

Mittet R. Second-order time integration of the wave equation with dispersion correction procedures // Geophysics. 2019. Vol. 84 (4). P. T221–T235. doi:10.1190/geo2018-0770.1.

Moczo P., Kristek J., Vavrycuk V., Archuleta R.J., Halada L. 3D heterogeneous staggered-grid finite-difference modeling of seismic motion with volume harmonic and arithmetic averaging of elastic moduli and

densities // Bulletin of the Seismological Society of America. 2002. Vol. 92 (8). P. 3042–3066. doi:10.1785/0120010167.

Siahkoohi A., Louboutin M., Herrmann F.J. The importance of transfer learning in seismic modeling and imaging // Geophysics. 2019. Vol. 84 (6). P. A47–A52. doi: 0.1190/geo2019-0056.1.

Virieux J. *P-SV* wave propagation in heterogeneous media: Velocity-stress finite-difference method // Geophysics. 1986. Vol. 51 (4). P. 889–901. doi:10.1190/1.1442147.

Virieux J., Calandra H., Plessix R.-E. A review of the spectral, pseudo-spectral, finite-difference and finite-element modelling techniques for geophysical imaging // Geophysical Prospecting. 2011. Vol. 59 (5). P. 794–813. doi:10.1111/j.1365-2478.2011.00967.x.

Vishnevsky D., Lisitsa V., Tcheverda V., Reshetova G. Numerical study of the interface errors of finite-difference simulations of seismic waves // Geophysics. 2014. Vol. 79 (4). P. T219–T232. doi:10.1190/geo2013-0299.1.

ИНФОРМАЦИЯ ОБ АВТОРАХ

ГАДЫЛЬШИНА Ксения Александровна – младший научный сотрудник лаборатории вычислительной физики горных пород Института нефтегазовой геологии и геофизики СО РАН. Основные научные интересы: методы машинного обучения в приложении к решению задач геофизики, https://orcid.org/0000-0003-0581-7741.

ЛИСИЦА Вадим Викторович – доктор физико-математических наук, заведующий лабораторией вычислительной физики горных пород Института нефтегазовой геологии и геофизики СО РАН. Основные научные интересы: численные методы для моделирования физических процессов в пористых средах.

ГАДЫЛЬШИН Кирилл Геннадьевич – кандидат физико-математических наук, старший научный сотрудник лаборатории вычислительной физики горных пород Института нефтегазовой геологии и геофизики СО РАН. Основные научные интересы: прямые и обратные задачи сейсмики, применение методов машинного обучения для повышения качества сейсмических данных, https://orcid.org/0000-0001-7247-6911.

ВИШНЕВСКИЙ Дмитрий Михайлович – научный сотрудник лаборатории вычислительной физики горных пород Института нефтегазовой геологии и геофизики СО РАН. Основные научные интересы: численное моделирование сейсмических волновых полей, высокопроизводительные вычисления, https://orcid.org/0000-0002-1439-4552.

КОСТИН Виктор Иванович – кандидат физико-математических наук, старший научный сотрудник лаборатории вычислительной физики горных пород Института нефтегазовой геологии и геофизики СО РАН. Основные научные интересы: численное моделирование сейсмических волновых полей, высокопроизводительные вычисления.

Статья поступила в редакцию 14 ноября 2023 г.,
одобрена после рецензирования 15 января 2024 г.,
принята к публикации 31 января 2024 г.