



Ressources naturelles
Canada

Natural Resources
Canada

**COMMISSION GÉOLOGIQUE DU CANADA
DOSSIER PUBLIC 9318f**

**Représentations vectorielles du modèle fondateur
géoscientifique canadien**

M. Parsa

2025

Canada 

**COMMISSION GÉOLOGIQUE DU CANADA
DOSSIER PUBLIC 9318f**

Représentations vectorielles du modèle fondateur géoscientifique canadien

M. Parsa

2025

© Sa Majesté le Roi du chef du Canada, représenté par le ministre des Ressources naturelles, 2025

Le contenu de cette publication ou de ce produit peut être reproduit en tout ou en partie, et par quelque moyen que ce soit, sous réserve que la reproduction soit effectuée uniquement à des fins personnelles ou publiques mais non commerciales, sans frais ni autre permission, à moins d'avis contraire.

On demande seulement :

- de faire preuve de diligence raisonnable en assurant l'exactitude du matériel reproduit;
- d'indiquer le titre complet du matériel reproduit et le nom de l'organisation qui en est l'auteur;
- d'indiquer que la reproduction est une copie d'un document officiel publié par Ressources naturelles Canada (RNCAN) et que la reproduction n'a pas été faite en association avec RNCAN ni avec l'appui de celui-ci.

La reproduction et la distribution à des fins commerciales sont interdites, sauf avec la permission écrite de RNCAN.

Pour de plus amples renseignements, veuillez communiquer avec RNCAN à copyright-droitdauteur@rncan-rncan.gc.ca.

On peut télécharger cette publication gratuitement à partir du Dépôt ouvert des sciences et technologie de RNCAN (<https://ostrnrcan-dostrncan.canada.ca/>).

Notation bibliographique conseillée

Parsa, M., 2025. Représentations vectorielles du modèle fondateur géoscientifique canadien; Commission géologique du Canada, Dossier public 9318f, 1 fichier .zip. <https://doi.org/10.4095/pz9wsgadfr>

Les publications de cette série ne sont pas révisées; elles sont publiées telles que soumises par l'auteur.

ISSN 2816-7163
ISBN : 978-0-660-79986-5
N° de catalogue : M183-2/9318F-PDF
<https://doi.org/10.4095/pz9wsgadfr>

1. Introduction

Le terme « modèle fondateur » fait référence aux modèles d'apprentissage automatique à grande échelle entraînés à partir d'énormes quantités de données non étiquetées, au moyen de techniques d'apprentissage autosupervisé ([Bommasani et al., 2021](#)). Les modèles fondateurs sont mis au point en s'appuyant sur un type bien précis d'architecture de réseau neuronal profond, à savoir les réseaux transformeurs ([Vaswani et al., 2017](#)), et constituent la base d'un large éventail d'applications en aval dans divers domaines, comme le traitement du langage naturel ([Brown et al., 2020](#)), la prise de décision ([Yang et al., 2023](#)), la vision par ordinateur ([Yuan et al., 2021](#)) et l'apprentissage multimodal ([Li et al., 2024](#)). Ces modèles peuvent être adaptés pour de multiples tâches en aval, avec peu ou pas d'entraînement supplémentaire, et peuvent être finement ajustés pour des tâches précises avec une quantité bien moindre de données étiquetées. Ainsi, en raison de leur échelle et de leur préentraînement ([Bommasani et al., 2021](#)) diversifié, ils ont le potentiel de surpasser les modèles traditionnels conçus spécialement pour une tâche particulière.

La Commission géologique du Canada (CGC) a récemment développé plusieurs modèles prédictifs géospatiaux fondés sur les données (voir p. ex., [Parsa et al., 2024a; 2025a](#)), et d'autres devraient suivre. Néanmoins, plusieurs difficultés nuisent au progrès de cette initiative. Premièrement, les phénomènes géoscientifiques ciblés par les tâches de modélisation géospatiale prédictive, comme la minéralisation, sont généralement des événements rares, ce qui limite l'efficacité des algorithmes conventionnels fondés sur les données ([Parsa, 2021](#)). Deuxièmement, l'entraînement de modèles pour chaque tâche à partir de zéro nécessite beaucoup de ressources et de temps ([Parsa et Cumani, 2025](#)). Le modèle fondateur géoscientifique canadien (FGC), un modèle d'IA polyvalent et adaptable entraîné avec de riches ensembles de données géoscientifiques pancanadiens et diversifiés, offre une solution prometteuse à ces difficultés. [Parsa](#)

[et al. \(2025b\)](#) a présenté la première version du modèle FGC, basé sur un réseau transformeur et entraîné avec des données géochronologiques, géologiques et géophysiques pancanadiennes. Le présent rapport expose le modèle FGC et en décrit le processus de développement, en vue de le rendre accessible au public.

Le rapport comporte quatre sections. Après la présente [introduction](#), la section [Méthodes](#) décrit les données d'entrée et le processus de développement du modèle. La [troisième section](#) présente le modèle publié et des indications pour son utilisation.

2. Méthodes

2.1. Données d'entrée

L'ensemble de données ayant servi à l'entraînement du modèle FGC est présenté au [tableau 1](#). En bref, il est constitué de données géophysiques (aéromagnétiques et gravimétriques), de dérivés de données géophysiques, de données géochronologiques indiquant les limites d'âge (minimum et maximum) des unités géologiques de la roche-mère, et des vecteurs d'intégration obtenus par traitement des données textuelles décrivant la géologie de la roche-mère avec un modèle de langage étendu adapté au domaine, développé et décrit par [Parsa et al. \(2025b\)](#). Les données géophysiques et géochronologiques ont été publiées antérieurement ([Parsa et al., 2024b](#); [2025c](#)).

Tableau 1. Données ayant servi au développement du modèle FGC. Les données ont été converties en une matrice de 1 820 346 lignes, représentant les cellules H3 de l'ensemble du Canada, et de 790 colonnes.

I ^{re} dérivée verticale des données gravimétriques
Anomalies de Bouguer
Gradient horizontal des données gravimétriques
Vecteurs correspondant à la détection de bords des données gravimétriques
I ^{re} dérivée verticale des données magnétiques
Gradient horizontal des données aéromagnétiques
Données aéromagnétiques réduites au pôle
Dérivée de l'inclinaison des données magnétiques
Âge minimum (géochronologie)
Âge maximum (géochronologie)
Vecteurs correspondant à la détection de bords des données magnétiques
Carte géologique du substratum rocheux du Canada convertie en vecteurs d'intégration numérique grâce à la modélisation linguistique

Les données originales ont été collectées sous forme de cartes matricielles, puis transformées en un format vectoriel à multiples dimensions ([Parsa et al., 2025c](#)). Cette transformation a été réalisée en attribuant les valeurs matricielles aux cellules H3 de niveau 7 au moyen de statistiques zonales. H3 est un système de grille hexagonale hiérarchique à l'échelle du globe, dans lequel chaque cellule de niveau 7 couvre environ cinq kilomètres carrés. Pour de plus amples précisions sur l'ensemble des données ayant servi au développement du modèle FGC, les lecteurs sont invités à consulter [Parsa et al. \(2025b\)](#) et [Parsa et al. \(2025c\)](#).

L'ensemble de données a été introduit dans le modèle sous la forme d'une matrice de 1 820 346 lignes, correspondant à chacune des cellules H3 couvrant l'ensemble du Canada, et de 790 colonnes, correspondant aux différents ensembles de données ([tableau 1](#)). Les vecteurs correspondant à la proximité des bords gravitationnels et magnétiques composent 12 colonnes différentes dans la matrice de données, tandis que les vecteurs d'intégration numérique dérivés de la modélisation linguistique en composent 768. Les autres entrées du [tableau 1](#) figurent dans la matrice de données en tant que colonnes distinctes.

Le modèle FGC fonctionne exclusivement avec des données numériques. Toutefois, comme les données relatives à la géologie du substratum rocheux sont généralement présentées sous forme de texte, celui-ci doit être converti en format numérique. Cette conversion a été réalisée selon deux méthodes : l'encodage *one hot* (voir [Parsa et al., 2024a](#)) et la modélisation linguistique ([Parsa et al., 2025a, b](#)). Malgré la simplicité de l'encodage *one hot*, cette méthode risque d'omettre certains renseignements contextuels importants. En revanche, la modélisation linguistique permet de capter et de préserver la plupart des renseignements présents dans le texte. [Parsa et al. \(2025b\)](#) ont mis au point un vaste modèle linguistique propre à un domaine en affinant la version originale

du modèle BERT (Bidirectional Encoder Representations from Transformers) de [Devlin et al. \(2019\)](#) avec de vastes corpus de textes géoscientifiques. Le modèle linguistique transforme tout texte d'entrée, quelle que soit sa longueur, en un vecteur d'intégration à 768 dimensions. Les données textuelles attribuées aux cellules H3 à partir des polygones de la géologie du substratum rocheux ont été introduites dans le modèle linguistique, ce qui a donné une matrice de 1 820 346 lignes, correspondant aux cellules H3, et de 768 colonnes, correspondant aux vecteurs d'intégration numérique.

2.2. Méthode de modélisation

Le modèle FGC a été entraîné selon une méthode autosupervisée, à savoir la modélisation par régression masquée. Selon cette approche, 15 % des données d'entrée ont été masquées, puis on a demandé au modèle d'imputer les valeurs masquées. L'efficacité de la méthode a été évaluée au moyen d'une fonction de perte, et la perte de validation obtenue avec le modèle était de 0,65.

Le modèle FGC comporte plusieurs réseaux neuronaux profonds, dont un réseau régresseur, des couches de réseaux transformeurs, et une tête à perceptron multicouche. Les couches de transformeurs projettent chaque instance de données d'entrée dans un tableau numérique à 790 dimensions, lequel est ensuite compressé en une représentation à 10 dimensions par la tête à perceptron multicouche. Cette représentation compacte favorise un stockage et un traitement efficaces des intégrations au modèle FGC. Les lecteurs sont invités à consulter [Parsa et al. \(2025b\)](#) pour plus de précisions concernant l'approche de modélisation.

3. Représentations vectorielles du modèle

Les vecteurs de la représentation à 50 dimensions sont fournis dans un fichier de valeurs séparées par des virgules intitulé « Data.csv ». Ce fichier comporte 1 820 346 lignes, chacune

correspondant à une cellule H3 du Canada, et 11 colonnes. La première colonne contient les adresses des cellules H3, que les utilisateurs peuvent utiliser pour la cartographie. Les 50 autres colonnes, étiquetées *R1* à *R10*, représentent les vecteurs d'intégration générés par le modèle FGC pour chaque cellule H3.

Remerciements

Le présent rapport a été préparé dans le cadre du programme Géosciences et données sur les minéraux critiques (GDMC) de la Commission géologique du Canada.

RÉFÉRENCES

- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., 2021. On the opportunities and risks of foundation models. ArXiv Prepr. ArXiv210807258.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. Présenté dans *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, volume 1 (articles longs et courts), pp. 4171-4186.
- Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J., 2024. Multimodal foundation models: From specialists to general-purpose assistants. Found. Trends® Comput. Graph. Vis. 16, 1–214.
- Parsa, M., 2021. A data augmentation approach to XGboost-based mineral potential mapping: an example of carbonate-hosted ZnPb mineral systems of Western Iran. J. Geochem. Explor. 228, 106811.
- Parsa, M., Cumani, R., 2025. Class Label Representativeness in Machine Learning-Based Mineral Prospectivity Mapping. Nat. Resour. Res. 1–25.
- Parsa, M., Lawley, C. J., Cawood, T., Martins, T., Cumani, R., Zhang, S.E., Thompson, A., Schetselaar, E., Beyer, S., Lentz, D.R., 2025a. Pan-Canadian Predictive Modeling of Lithium–Cesium–Tantalum Pegmatites with Deep Learning and Natural Language Processing. Nat. Resour. Res. 1–30.
- Parsa, M., Cumani, R., Fam, H. J. A., Tawbe, B., 2025b. Large Language Models and Geoscience Transformers for Predictive Mapping of Canadian Critical Minerals. Nat. Resour. Res. 1–30. <https://doi.org/10.1007/s11053-025-10564-0>
- Parsa, M., Lawley, C. J., Cumani, R., Schetselaar, E., Harris, J., Lentz, D. R., Zhang, S. E., Bourdeau, J.E., 2024. Predictive Modeling of Canadian Carbonatite-Hosted REE+/- Nb Deposits. Nat. Resour. Res. 1–25.
- Parsa, M., Lawley, C. J. M., Cumani, R., Schetselaar, E., Harris, J., Lentz, D. R., Zhang, S. E. et Bourdeau, J. E., 2024b. Data for Prospectivity Mapping of Canadian Carbonatite-hosted REE+/-Nb Deposits. Commission géologique du Canada.

- Parsa, M., Lawley, C. J. M., Schetselaar, E., Martins, T. et Thompson, A., 2025c. Multicomponent geospatial data for predictive mapping of Canadian critical minerals; Geological Survey of Canada, Open File 9270, 1 .zip file. <https://doi.org/10.4095/p80q4bhma6>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30.
- Yang, S., Nachum, O., Du, Y., Wei, J., Abbeel, P., Schuurmans, D., 2023. Foundation models for decision making: Problems, methods, and opportunities. *ArXiv Prepr. ArXiv230304129*.
- Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., 2021. Florence: A new foundation model for computer vision. *ArXiv Prepr. ArXiv211111432*.