Implementing GeoAI with KNIME and its Geospatial Analytics Extension

Ru Wang

Nov. 7th, 2023

GeoAI & Social Science

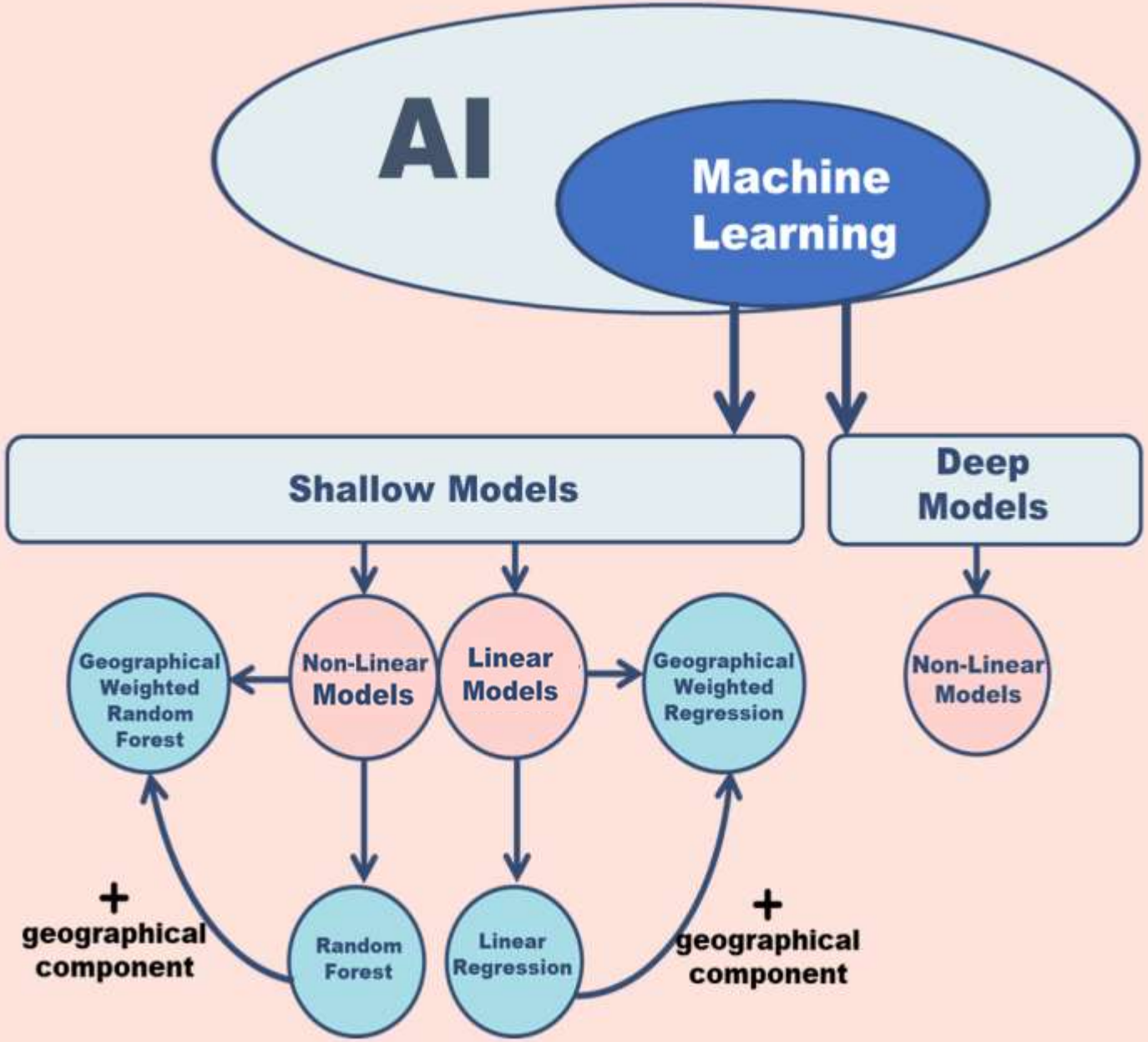
Social Science & GeoAI

- In many **social science studies, spatial relationships** between *human communities*, *cultures*, *economies*, and their interactions with the environment are always being required to explored.

- **Regression analysis** is a powerful statistical method widely used to determine the relationship between one or more independent variables (explanatory variables) and a dependent variable (response variable).

- **AI/GeoAI** provide us with various options for regression analysis research.

Copyright by Spatial Data Lab @ Center for Geographic Analysis, Harvard University

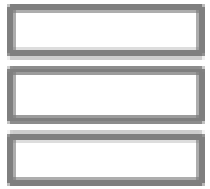


What is GeoAI ?

Geospatial artificial intelligence (GeoAI) is an emerging and promising research field that integrates AI with geospatial science to resolve problems and issues of geographic nature (Li and Hsu, 2022).

Algorithm to Be Investigated

1.Linear Regression



$$Y_i = ax_i + e, i = 1 : n$$

ax_i: **linear** prediction

2.Geographical Weighted Regression

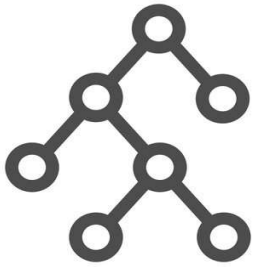
$$Y_i = ax_i + e, i = 1 : n$$

linear prediction

Copyright by Spatial Data Lab @ Center for Geographic Analysis, Harvard University

Algorithm to Be Investigated

3.Random forest



$$Y_i = ax_i + e, i = 1 : n$$

ax_i: **nonlinear** prediction

4.Geographical Random forest

$$Y_i = ax_i + e, i = 1 : n$$

ax_i: **nonlinear** prediction

Copyright by Spatial Data Lab @ Center for Geographic Analysis, Harvard University

How to Interpret RF?

A Unified Approach to Interpreting Model Predictions

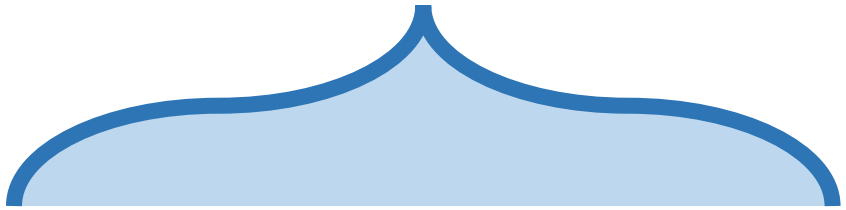
Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

SHAP



SHapley **A**dditive ex**P**lanation

A game theoretic approach **to explain the** output of any machine learning model. **It connects optimal credit allocation with local explanations using the classic Shap values from game theory and their related extensions.**

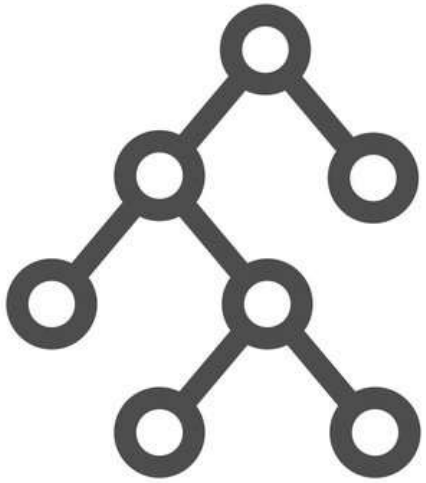
Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Copyright by Spatial Data Lab @ Center for Geographic Analysis, Harvard University

How to Interpret RF?

SHAP

Interpret



Random forest

Copyright by Spatial Data Lab @ Center for Geographic Analysis, Harvard University

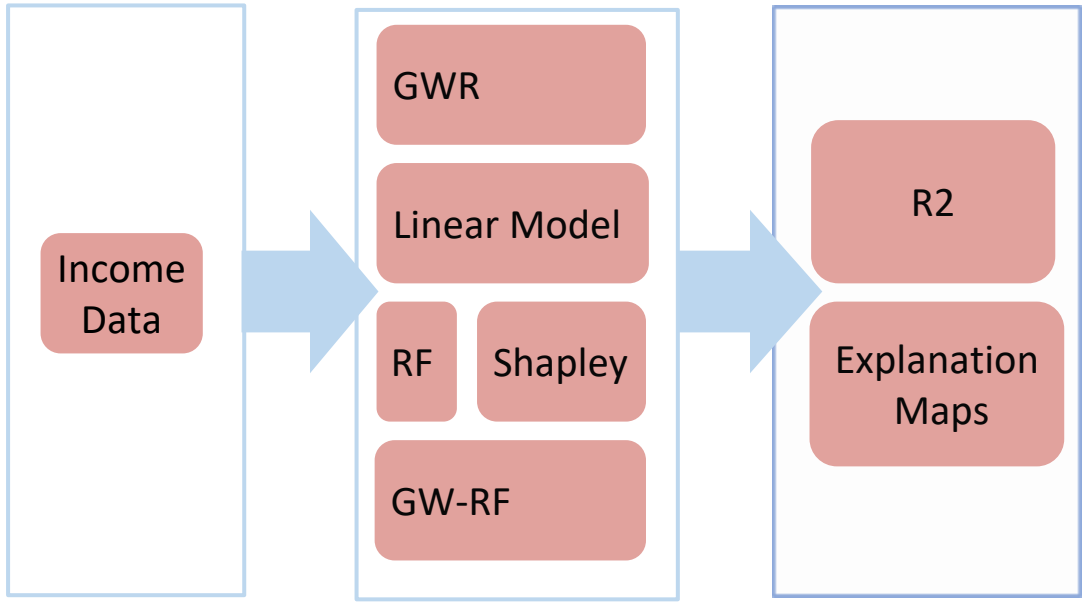
Case Studies with KNIME

Predicting Household Income in Greece

A Comparative Analysis of Linear Regression, Geographically Weighted Regression, Random Forest, and Geographically Weighted Random Forest Models.

Predicting Household Income in Greece

Workflow overview:



Overview:

✓ **Build a workflow** to predict the household income in Greece (Programme Kallikratis) using different models including linear regression, geographically weighted regression, random forest, and Geographically Weighted Random Forest in Knime.

✓ **Use Shap to explain the prediction results** of the random forest.

✓ **Compare the results** of different models and understand the advantages and disadvantages of different AI/GeoAI models.

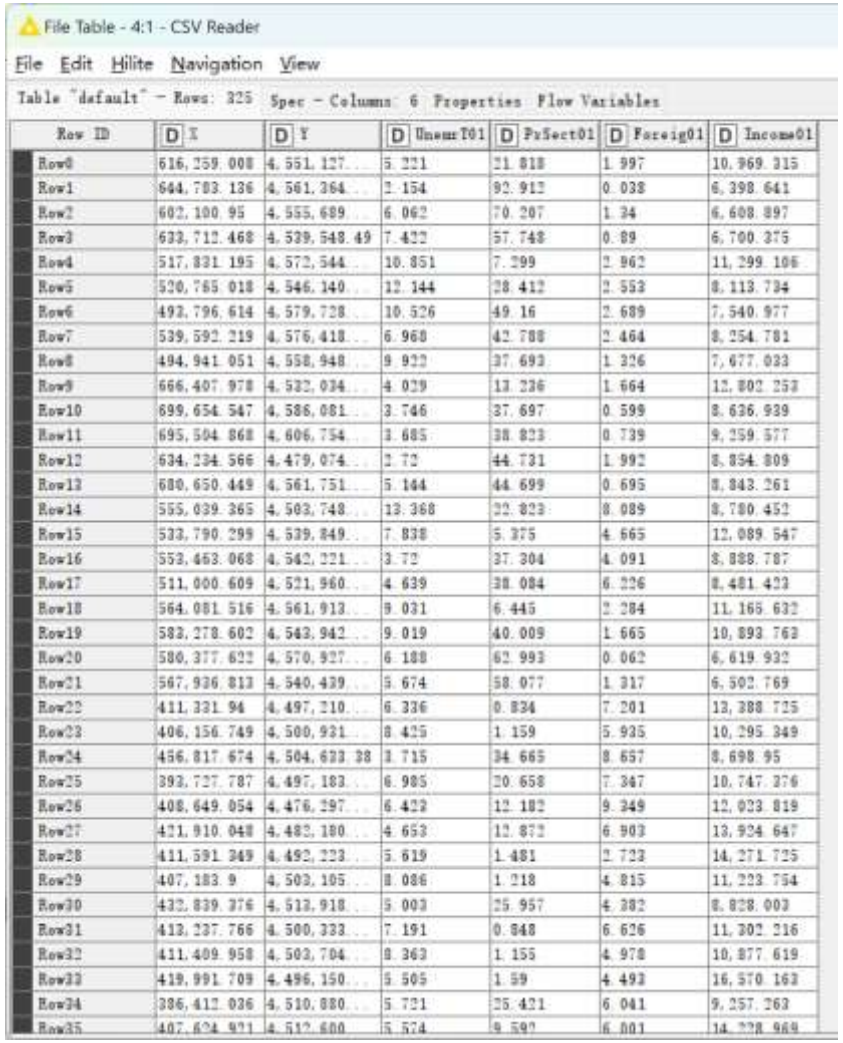
Copyright by Spatial Data Lab @ Center for Geographic Analysis, Harvard University

Predicting Household Income in Greece

Input data:

The data used in this tutorial is from the R SpatialML package. You can download from the URL. This data have 325 observations on the following variables.

- **X:** x coordinates in EPSG:2100
- **Y:** y coordinates in EPSG:2100
- **UnemrT01(Independent Variable):** the total unemployment rate in 2001 (Census)
- **PrSect01(Independent Variable):** the proportion of economically active working in the primary financial sector (mainly agriculture; fishery; and forestry in 2001 (Census))
- **Income01(Dependent Variable):** mean recorded household income (in Euros) earned in 2001 and declared in 2002 tax forms



File Table - 4:1 - CSV Reader

File Edit Hilite Navigation View

Table "default" – Rows: 325 Spec – Columns: 6 Properties Flow Variables

Row ID	X	Y	UnemrT01	PrSect01	Foreig01	Income01
Row0	616, 259. 008	4, 551, 127.	5. 221	21. 818	1. 997	10, 969. 315
Row1	644, 783. 136	4, 561, 364.	2. 154	92. 912	0. 038	6, 398. 641
Row2	602, 100. 95	4, 555, 689.	6. 062	70. 287	1. 34	6, 608. 897
Row3	633, 712. 468	4, 539, 548. 49	7. 422	57. 748	0. 89	6, 700. 375
Row4	517, 831. 195	4, 572, 544.	10. 851	7. 299	2. 962	11, 299. 106
Row5	520, 765. 018	4, 546, 140.	12. 144	28. 412	2. 553	8, 113. 734
Row6	493, 796. 614	4, 579, 728.	10. 526	49. 16	2. 689	7, 540. 977
Row7	539, 592. 219	4, 576, 418.	6. 968	42. 788	2. 464	8, 254. 781
Row8	494, 941. 051	4, 558, 948.	9. 922	37. 693	1. 326	7, 677. 033
Row9	666, 407. 978	4, 532, 034.	4. 029	13. 236	1. 664	12, 802. 253
Row10	699, 654. 547	4, 586, 081.	3. 746	37. 697	0. 599	8, 636. 939
Row11	695, 504. 868	4, 606, 754.	3. 685	38. 823	0. 739	9, 359. 577
Row12	634, 234. 566	4, 479, 074.	2. 72	44. 731	1. 992	8, 854. 809
Row13	680, 650. 449	4, 561, 751.	5. 144	44. 699	0. 695	8, 843. 261
Row14	555, 039. 365	4, 503, 748.	13. 368	22. 823	8. 089	8, 780. 452
Row15	533, 790. 299	4, 539, 849.	7. 838	5. 375	4. 665	12, 089. 547
Row16	553, 463. 068	4, 542, 221.	3. 72	37. 304	4. 091	8, 888. 787
Row17	511, 000. 609	4, 521, 960.	4. 639	38. 084	6. 226	8, 481. 423
Row18	564, 081. 516	4, 561, 913.	9. 031	6. 445	2. 284	11, 165. 632
Row19	583, 278. 602	4, 543, 942.	9. 019	40. 009	1. 665	10, 893. 763
Row20	580, 377. 622	4, 570, 927.	6. 188	62. 993	0. 062	6, 619. 932
Row21	567, 936. 813	4, 540, 439.	5. 674	58. 077	1. 317	6, 502. 769
Row22	411, 331. 94	4, 497, 210.	6. 336	0. 834	7. 201	13, 388. 725
Row23	406, 156. 749	4, 500, 931.	8. 425	1. 159	5. 935	10, 295. 349
Row24	456, 817. 674	4, 504, 623. 38	3. 715	34. 665	8. 657	8, 698. 95
Row25	393, 727. 787	4, 497, 183.	6. 985	20. 658	7. 347	10, 747. 376
Row26	408, 649. 054	4, 476, 297.	6. 422	12. 182	9. 349	12, 023. 819
Row27	421, 910. 048	4, 483, 180.	4. 653	12. 872	6. 903	13, 924. 647
Row28	411, 591. 349	4, 492, 223.	5. 619	1. 481	2. 723	14, 271. 725
Row29	407, 183. 9	4, 503, 105.	8. 086	1. 218	4. 815	11, 223. 754
Row30	432, 839. 376	4, 513, 918.	5. 003	25. 957	4. 382	8, 828. 003
Row31	413, 237. 766	4, 500, 333.	7. 191	0. 848	6. 626	11, 302. 216
Row32	411, 409. 958	4, 502, 704.	8. 363	1. 155	4. 978	10, 877. 619
Row33	419, 991. 709	4, 496, 150.	5. 505	1. 59	4. 493	16, 570. 163
Row34	386, 412. 036	4, 510, 880.	5. 721	25. 421	6. 041	9, 257. 263
Row35	407, 624. 971	4, 512, 600.	5. 574	9. 592	6. 801	14, 228. 969

https://cran.r-project.org/web/packages/SpatialML/index.html

Copyright by Spatial Data Lab @ Center for Geographic Analysis, Harvard University

Predicting Household Income in Greece

SpatialML: Spatial Machine Learning R package

Implements a spatial extension of the random forest algorithm <https://www.mdpi.com/2220-9964/11/9/471>.

Install:

- ✓ Install the R.
- ✓ Install the KNIME Interactive R Statistics Integration. Please refer to this guide for more information about the installation.
- ✓ Install the SpatialML R package. Open the R shell and install the SpatialML package by running the command install.packages("SpatialML").

Usage: grf (formula, dframe, bw, kernel, coords, ntree=500, mtry=NULL, importance="impurity", nthreads = NULL, forests = TRUE, weighted = TRUE, print.results=TRUE, ...)

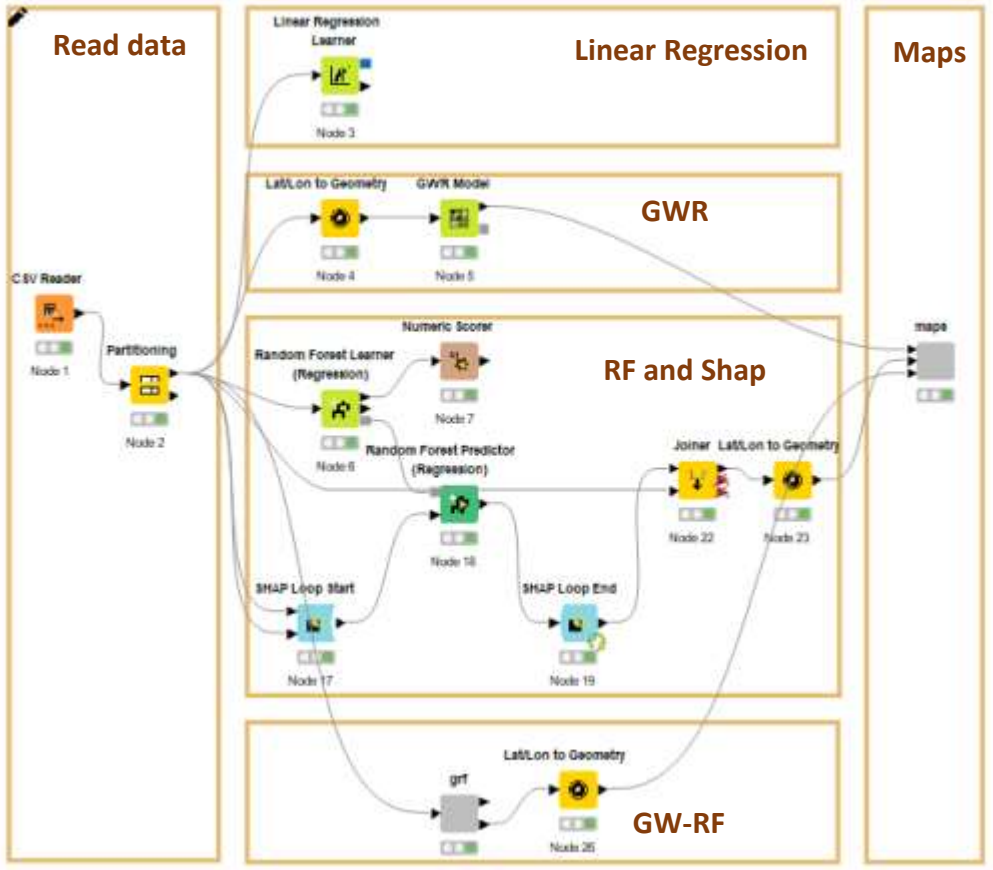
Example:

```
1  library(SpatialML)     # Geographically weigted regression
2  data(Income)
3  Coords<-Income[ ,1:2]
4  grf <- grf(Income01 ~ UnemrT01 + PrSect01, dframe=Income, bw=60,
5  kernel="adaptive", coords=Coords)
```

Copyright by Spatial Data Lab @ Center for Geographic Analysis, Harvard University

Predicting Household Income in Greece

Workflow in KNIME:

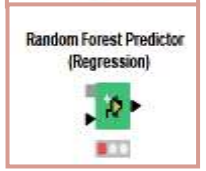


Nodes in workflow:

Performs a multivariate linear regression.

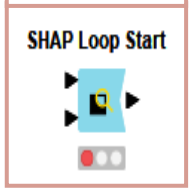
Performs Geographically Weighted Regression.

Learns a random forest* for regression.

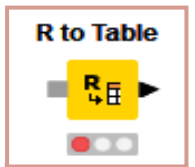


Applies regression from a random forest* model.

Calculates the SHAP values.



SHAP represents a unified approach to explain the predictions of any machine learning model.



Converts a data frame from an R workspace into a KNIME data table.

Converts a KNIME data table into a data frame which is accessible within R as knime.in.

The data table from the input port is imported into the R workspace as variable knime.in;existing variables are overwritten in case of a conflict.

Copyright by Spatial Data Lab @ Center for Geographic Analysis, Harvard University

Predicting Household Income in Greece

Results:

- **The linear regression model** is easy to understand and explainable, however, our data has a strong spatial effect, so the linear regression model is not a good model for our data.
- **The GWR model** is a good model for our data since it considers the spatial effect.
- **The random forest model** is a non-linear model, so it is better than the linear regression model but due to the spatial effect, it is not as good as the GWR model.
- **The GW-RF model** is the best model among all the models since it considers the spatial effect and it is a non-linear model.

Model	R2
Linear Regression	0.627
GWR	0.842
Random Forest	0.686
GW-RF	0.98

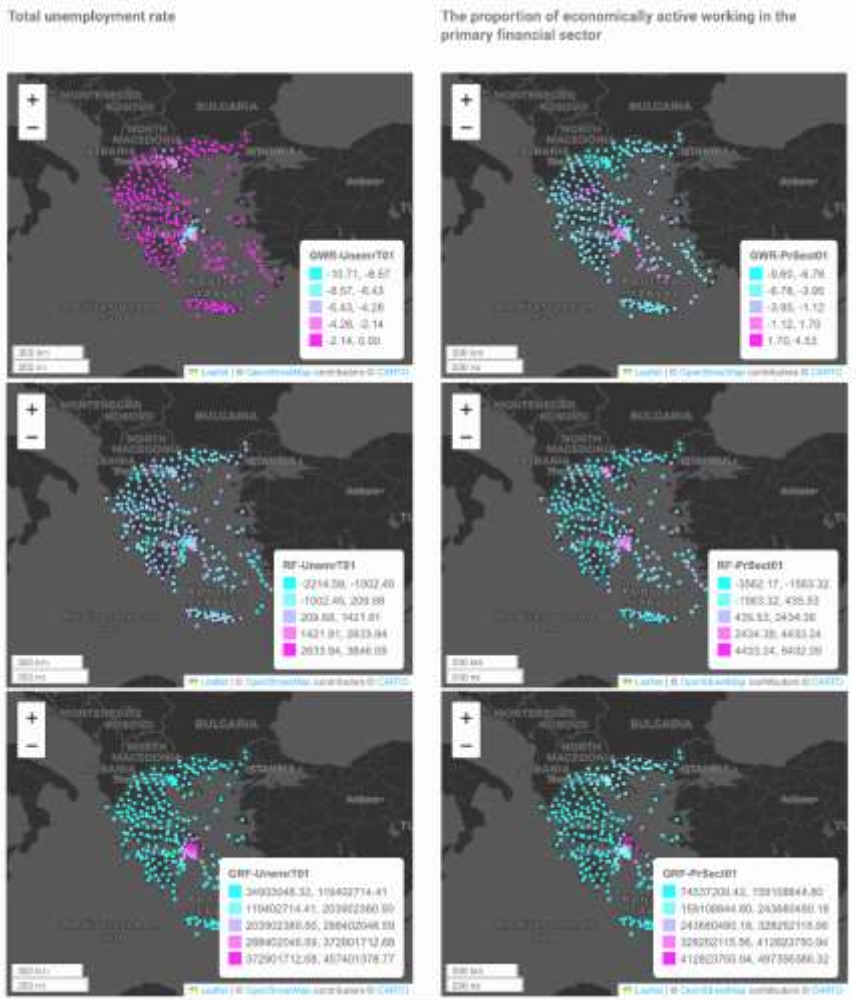
Copyright by Spatial Data Lab @ Center for Geographic Analysis, Harvard University

Predicting Household Income in Greece

Results:

- We also explored the **spatial effects** of the total unemployment rate and the proportion of economically active working in the primary financial sector on predicting household income in Greece.

- We found that the total unemployment rate and the proportion of economically active working in the primary financial sector contribute more in the center of the map and less in the edge of Greece.



Copyright by Spatial Data Lab @ Center for Geographic Analysis, Harvard University

Conclusion

- ✓ Learn **some background in regression analysis model of AI/Geo-AI**.
- ✓ Build a workflow to **implement different models** including linear regression, geographically weighted regression, random forest, and Geographically Weighted Random Forest in KNIME.
- ✓ Write **R code** to implement the geographically weighted random forest and **integrate the R code** into Knime.
- ✓ Use **Shap to explain the prediction results** of any machine learning or AI models in Knime.
- ✓ **KNIME** provide us a platform to easily and visually **compare the results of different regression models** and understand the advantages and disadvantages of different models.

Copyright by Spatial Data Lab @ Center for Geographic Analysis, Harvard University

Implementing
GeoAI with KNIME and its Geospatial Analytics Extension