Data4Resilience: Challenge and Database Overview

Leveraging Climate and Demographic Data for Resilience Research

MEMOREC Staff

University of Bologna

January 14, 2025

Introduction to the Challenge

- A data challenge by MEMOREC to explore innovative uses of open data.
- Focused on climate adaptation and demographic resilience.
- Open to students, researchers, and data enthusiasts.

Questions and FAQ

How to Ask Questions:

- Please write your questions in the chat during the session.
- We will collect all questions and address them at the end of the workshop.
- After the event, we will compile and share a document with Frequently Asked Questions (FAQ) for all participants.

Your participation is valuable!

Philosophy of the MEMOREC Project

- **Interdisciplinary Approach:** Integrates demographic, climatic, and geographic data.
- **Open-Access Database:** Useful for studying the impact of climate change on Emilia-Romagna.
- **Spatiotemporal Analysis:** Maps sociodemographic vulnerabilities related to extreme climatic events.
- **Historical Documentation:** Enhances historical photographic records and reconstructs statistical sources.
- **Community Resilience:** Explores how communities have adapted to climatic impacts over time.
- **Policy Contribution:** Aims to inform effective territorial planning and policies.

Objectives of the Challenge

- Encourage creative analyses using MEMOREC data.
- Foster interdisciplinary collaboration.
- Highlight the importance of resilience in the face of climate change.
- The goal of the Data Challenge is to test potential issues in the beta version of the database in preparation for its final release.

Overview of the MEMOREC Database - EcoPopER Beta Version

- A collection of historical and contemporary datasets.
- Covers climate, demographic, and environmental data for Emilia-Romagna.
- Supports resilience research and innovation.

Key Features of the Database

- **Interdisciplinary:** Combines climate, demographic, and environmental data.
- **Historical Scope:** From 2000 to the present.
- **Open Access:** Freely available.

EcoPopER Open Data: FAIR Principles

The EcoPopER database adheres to FAIR principles:

- **Findable:** Data are easily discoverable through standardized metadata.
- **Accessible:** Data can be accessed using open and universally applicable protocols.
- **Interoperable:** Data are formatted to work seamlessly with other datasets and tools.
- **Reusable:** Data are well-documented and licensed for future use by the scientific community.

FAIR principles ensure that open data maximize their potential impact in research and policy.

Dataset Categories

- **Climate Data:** Temperature, precipitation, extreme weather events.
- **Demographic Data:** Population registers, vital statistics.
- **Environmental Data:** Land use, urbanization.
- **Resilience Indicators:** Socioeconomic vulnerability factors.

Dataset Structure

- The dataset is provided in a **long form** format.
- Each row represents a single observation (e.g., a municipality at a specific time point).
- Includes multiple variables such as demographic indicators, climate data, and geographic attributes.
- Easily transformable into a **wide form** format for comparative analysis or visualization.
- Suitable for time-series analysis and spatiotemporal modeling.

Dataset Access on Zenodo

Access and Use of the EcoPopER Dataset:

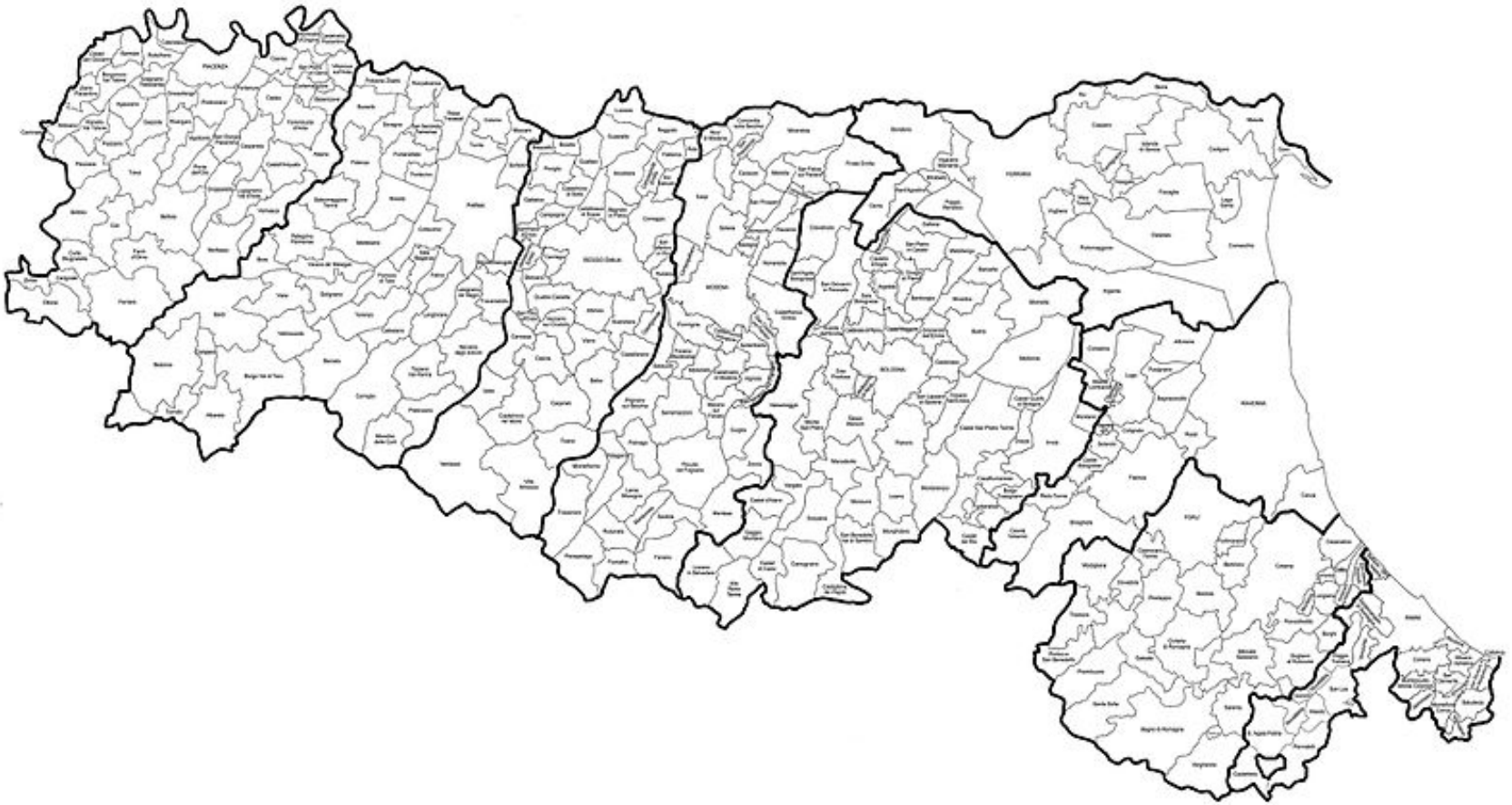
- The dataset is accessible on **Zenodo**: zenodo.org/records/13951348.
- Metadata files include detailed explanations about:
  - Variables and their definitions.
  - Coding schemes and modalities.
  - Instructions for using the dataset effectively.
- Ensure you review all metadata files for comprehensive understanding.

dataset_versione1_25102024.csv

Anno	Mese	Nome_Unità_Territoriale	Codice_Unità_Territoriale	Valore	Indicatore	Unità_
2004	1	AGAZZANO	33001	1	Decessi maschi	1
2004	2	AGAZZANO	33001	2	Decessi maschi	1
2004	3	AGAZZANO	33001	3	Decessi maschi	1

Territorial and Temporal Detail

- **Territorial Scope:** Covers all municipalities in the Emilia-Romagna region.
- **Temporal Coverage:** Spans approximately from 2000 to 2023.
- **Frequency:** Primarily available as monthly series.



Specifics of Database Indicators

- **Births, Deaths and Meteorological Indicators:** Cover monthly series from 2000 to 2023.
- **Other Indicators:**
  - May cover shorter time intervals.
  - May refer only to January 1 of certain years.
  - Examples: Geographic risk indicators, land use, condition of the built environment.

Dataset Challenges: Changing Boundaries

- In Emilia-Romagna, some municipalities have changed their territorial boundaries over the years covered by the dataset.
- Examples of changes:
  - Municipalities that transitioned from the Marche region to Emilia-Romagna.
  - Municipalities that merged into new entities.
- Some indicators are based on the current 2024 boundaries, while others refer to historical boundaries that vary over time.
- These changes must be considered when conducting analyses.

Examples of Indicators by Municipal Boundaries

Indicators for Municipalities with Fixed Boundaries Over Time:

- Temperature and Precipitation
- Deaths by Broad Age Groups (e.g., Elderly)
- Indicators of Social Vulnerability (e.g., Fragility Index)

Indicators for Municipalities with Changing Boundaries Over Time:

- Resident Population
- Monthly Deaths
- Monthly Births

Suggested Solutions for Boundary Changes

- To address these issues, we recommend:
  - Focusing on provinces where municipalities have not experienced boundary changes.
  - Municipalities formerly part of the Marche region may be excluded from the analysis.
  - Using the script provided to reconstruct indicators for municipalities based on their current boundaries.
- Additional documentation and support for handling these challenges are available on the MEMOREC website.

Scripts and Suggestions for Analysis

Facilitating Your Analysis with Prebuilt Scripts:

- We have developed scripts to:
  - Facilitate the selection of specific indicators.
  - Reconstruct administrative boundaries for analyses requiring consistent geographies.
- These tools are designed to reduce the complexity of your data processing tasks.
- **The scripts will be distributed today to all participants.**

Simplify Your Analysis:

- It is recommended to focus on a limited number of indicators.
- Prioritize indicators most relevant to your research question or hypothesis.

Example Research Questions

- What has been the impact of recent heatwaves on mortality, especially among the elderly?
- How much of the population lives in areas particularly at risk?
- What are the characteristics of the population living in high-risk areas?
- How has the population distribution changed in recent years?
- Which areas have been most affected by depopulation trends?
- Where is the most vulnerable population concentrated?
- What is the relationship between land use changes and demographic shifts?

Note: You do not need to answer all of these questions. Select only one, or propose your own research question.

Possible Suggested Analyses

- **Descriptive:**
  - Creation of maps to compare one or more indicators.
  - Derivation of new indicators.
  - Scatter plots between indicators and correlation indices.
  - Data visualization techniques and dashboard creation.
- **Multivariate Statistics:**
  - Cluster analysis.
  - PCA and data dimension reduction.
- **Statistical Modeling:**
  - Multiple regressions.
  - Time-series analysis.
  - Spatiotemporal models.
  - Count data regression (e.g., Poisson regressions for mortality).

Suggested Software for Analysis

- **Data Visualization and Mapping:**
  - QGIS for geographic mapping.
  - Tableau or Power BI for dashboards and data exploration.
- **Statistical Analysis:**
  - R and RStudio for advanced statistical and graphical analyses.
  - Python (libraries such as pandas, matplotlib, seaborn, statsmodels).
  - Stata for statistical modeling.
- **Machine Learning and Advanced Modeling:**
  - Scikit-learn or TensorFlow for predictive modeling.
  - MATLAB for multivariate and numerical modeling.

Using the Database for the Challenge

- Download datasets from the MEMOREC website.
- Analyze and visualize the data creatively.
- Submit findings by the deadline.
- **Don't forget to include the names of all group members in your submission.**

What to Submit for the Data Challenge

To participate in the Data Challenge, you must submit:

- **Results of Your Analysis:**
  - One or more items derived from the MEMOREC database.
- **Scripts Used:**
  - Provide any scripts or codes used for data processing and analysis.
- **Technical Report:**
  - A brief technical document explaining the steps followed during the analysis.
- **Discussion of Results:**
  - Comment on and interpret the results obtained from your analysis.

Where to Submit:

Send all materials to: `francesco.scalone@unibo.it`

Challenge Timeline

- Submission Deadline: 31 March 2025

Evaluation Criteria

- Creativity and originality.
- Relevance to resilience and climate adaptation.
- Methodological rigor.
- Quality of presentation.

Applications of MEMOREC Data

- Research on climate impacts on populations.
- Policy development for demographic resilience.
- Data visualization and storytelling projects.

Future Developments

- Upcoming Version 2.0 with additional datasets.
- Expanding the temporal and spatial coverage.

Prizes and Recognition

- Recognition by MEMOREC.
- Collaboration opportunities.
- Results featured on the MEMOREC platform.

R scripts suggestions

- Some suggestions to prepare your data before the analysis
- Overcome some difficulties
- You don't have to use this scripts or even R, these are just suggestions
- We will provide you with all the script presented here

How to select only certain indicators

I can select only some indicators to make my data more manageable and easier to work with.

You have to install the package tidyverse if you don't have it already.

```
library(tidyverse)
dati<-read.table("C:/Users/nbarb/OneDrive/Desktop/tabella/dataset_versioni/dataset_
  versione1_25102024.txt", header=T)
head(dati)
```

```
dati <- dati %>%
filter(grepl("^Decessi femmine tra|^Decessi maschi tra|^Precipitazioni|^Temperatura|^
  Decessi femmine con|^Decessi maschi con|^Popolazione residente totale di tutte|^
  Popolazione residente maschi di tutte|^Popolazione residente femmine di tutte",
  Indicatore))
unique(dati$Indicatore)
```

Note that the string in the grepl function should be on one line only .

Keep in mind that the script provided works for this selection of indicators but may not work for others.

How to align the municipalities with the present-day-ones

First, we need to change the ISTAT code for the municipalities that changed their ISTAT code during the past 25 years, for which more than one code is present in the dataset. Only Montecopiolo and Sassofeltrio have 2 different codes in the dataset.

```
result <- dati %>%
group_by(Nome_Unità_Territoriale) %>%
summarise(number_of_codes=n_distinct(Codice_Unità_Territoriale))
%>% filter(number_of_codes > 1)
```

```
# A tibble: 2 × 2
  Nome_Unità_Territoriale number_of_codes
  <chr>                            <int>
1 MONTECOPIOLO                         2
2 SASSOFELTRIO                         2
```

The actual change:

```
#montecopiolo e sassofeltrio
montecopiolo<-dati[dati$Nome_Unità_Territoriale=="MONTECOPIOLO",]
unique(montecopiolo$Codice_Unità_Territoriale)
check<-dati[dati$Codice_Unità_Territoriale==41033,]
dati$Codice_Unità_Territoriale<-ifelse(dati$Nome_Unità_Territoriale=="MONTECOPIOLO",
99030,dati$Codice_Unità_Territoriale)

sassofeltrio<-dati[dati$Nome_Unità_Territoriale=="SASSOFELTRIO",]
unique(sassofeltrio$Codice_Unità_Territoriale)
check<-dati[dati$Codice_Unità_Territoriale==41060,]
dati$Codice_Unità_Territoriale<-ifelse(dati$Nome_Unità_Territoriale=="SASSOFELTRIO",
99031,dati$Codice_Unità_Territoriale)

#province
dati$Provincia<-ifelse(dati$Nome_Unità_Territoriale=="SASSOFELTRIO","RIMINI",dati$
    Provincia)
dati$Provincia<-ifelse(dati$Nome_Unità_Territoriale=="MONTECOPIOLO","RIMINI",dati$
    Provincia)
```

We now need to reunite all the municipalities that were merged over the past 25 years. You can find a list of all the municipalities involved at: Codici delle unità amministrative.

Example (you need to do this for every new municipality):

```
#berra and ro are merged to create riva del po in 2019
dati$Nome_Unità_Territoriale<-ifelse(dati$Nome_Unità_Territoriale%in%c("BERRA","RO"),
"RIVA DEL PO",dati$Nome_Unità_Territoriale)
riva<-dati[dati$Nome_Unità_Territoriale=="RIVA DEL PO",]
unique(riva$Codice_Unità_Territoriale)
dati$Codice_Unità_Territoriale<-ifelse(dati$Codice_Unità_Territoriale%in%c(38002,38020)
    ,38029,dati$Codice_Unità_Territoriale)

#tresigallo amd formignana are merged to create tresignana in 2019
dati$Nome_Unità_Territoriale<-ifelse(dati$Nome_Unità_Territoriale%in%c("TRESIGALLO","
    FORMIGNANA"),"TRESIGNANA",dati$Nome_Unità_Territoriale)
tre<-dati[dati$Nome_Unità_Territoriale=="TRESIGNANA",]
unique(tre$Codice_Unità_Territoriale)
dati$Codice_Unità_Territoriale<-ifelse(dati$Codice_Unità_Territoriale%in%c(38009,38024)
    ,  38030,dati$Codice_Unità_Territoriale)
```

In the provided script the whole code is present.

Now there will be municipalities that appear twice (or even more times) for the same indicator and the same period: we need to collapse these values together.

Important: here I use the function *sum()* because population counts are the only indicators affected by the merge (the other indicators i.e. deaths and climatic indicators refer already to the current municipalities). If you have other indicators, you may want to use different functions such as *mean()* or compute a weighted mean (for percentages for example).

```
dati_aggregati <- dati %>%
group_by(Anno,
    Mese,
    Nome_Unità_Territoriale,
    Codice_Unità_Territoriale,
    Indicatore,
    Unità_Territoriale,
    Codice_Indicatore,
    Periodicità_Indicatore,
    Provincia) %>%
summarise(Valore_Totale = sum(Valore, na.rm = TRUE)) %>% #here sum
ungroup()
```

How to manage the municipalities that had been part of the Marche Region

Now that we have aligned the municipalities with the present-day-ones we should have 330 municipalities for each year, right? Except for certain years and for certain indicators, they are 321 or 328, why?
Problem: Some municipalities were transferred from the Marche Region to the Emilia-Romagna Region at some point during the past 25, so for certain indicators the data relative to these municipalities are missing.
At this point, we can either:

- Retrieve the data for the Marche Region (where possible) and integrate them (complicated);
- Remove these municipalities from the analysis.

The easiest option is to remove altogether these municipalities from the analysis.

```
#elimino dati relativi ai comuni che sono stati nelle marche ####

comuni_marche<-toupper(c("Casteldelci",
        "Maiolo",
        "Novafeltria",
        "Pennabilli",
        "San Leo",
        "Sant'Agata Feltria",
        "Talamello",
        "Montecopiolo",
        "Sassofeltrio"))
dati2<-ecopop%>%filter(!Nome_Unità_Territoriale%in%comuni_marche)
unique(dati2$Nome_Unità_Territoriale) #ora sono 330-9=321
```

Now the total number of municipalities is 321.

How to transform the dataset from long to wide format

```
dati2<-dati2%>%
select(!c(Codice_Indicatore,Periodicità_Indicatore))


dati_wide <- dati2 %>%
pivot_wider(
id_cols = c(Anno, Mese, Nome_Unità_Territoriale, Codice_Unità_Territoriale,
Provincia, Unità_Territoriale),
names_from = Indicatore,
values_from = Valore
)
```

How to create maps in R

First, we need to retrieve the shapefile containing the municipalities of the Emilia-Romagna Region in a certain year and join it with our dataset.
You can find them here: Confini delle unità amministrative a fini statistici.
You can use different packages to deal with shapefiles, one of the most used is the **sf** package.
Load the shapefile and select only the municipalities of Emilia-Romagna:

```
confini_com<-st_read("C:/Users/nbarb/OneDrive/Desktop/confini_amm/Limiti01012024_g/
   Limiti01012024_g/Limiti01012024_g/Com01012024_g/Com01012024_g_WGS84.shp")

#emilia romagna is code 8
confini_er<-confini_com[confini_com$COD_REG==8,]
```

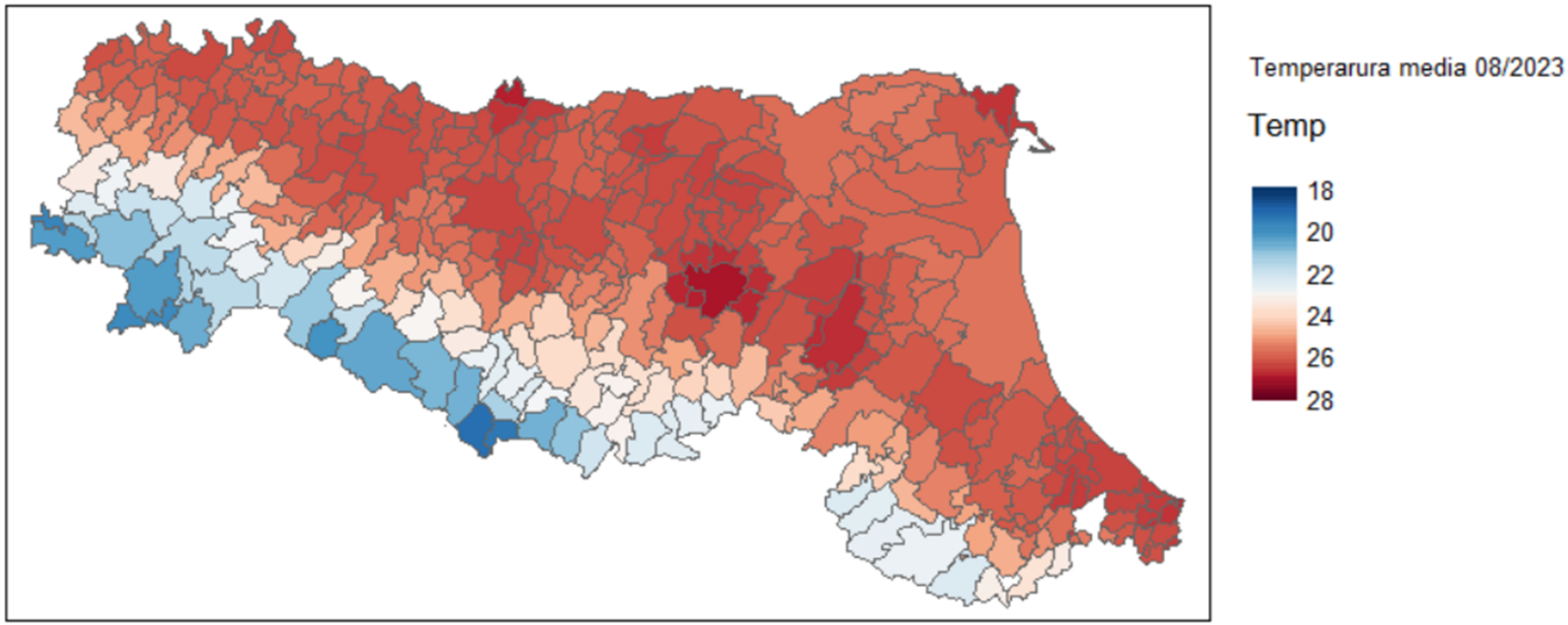
To join our data and the shapefile we can use either the ISTAT code or the municipality name (here I used the municipality name):

```
#I change the name of the column so that I can join the 2 dataframes
head(confini_er)
colnames(confini_er)[8]<-"Nome_Unità_Territoriale"
confini_er$Nome_Unità_Territoriale<-toupper(confini_er$Nome_Unità_Territoriale)


#I check that there aren't differences in the municipalities names of the 2 files
comuni_ecopop<-unique(dati$Nome_Unità_Territoriale)
setdiff(comuni_ecopop,confini_er$Nome_Unità_Territoriale)
setdiff(confini_er$Nome_Unità_Territoriale,comuni_ecopop)
#there aren't

join<-inner_join(confini_er, dati_wide, by="Nome_Unità_Territoriale")
```

You can use the tmap and ggplot2 packages to create maps. Example of what you can get:



Contact and Support

- Questions?
- Visit: https://site.unibo.it/memorec-resilienza-memoria-climatica
- Email: francesco.scalone@unibo.it