Radiation Exposure Analytics with Apache Spark

Stefani DW Yates

2025-04-22

Analysis of Radiation Exposure

Radiation surveillance is a public health imperative. Ionizing radiation, even at low doses, is linked to increased cardiovascular and circulatory disease risk—especially among occupational and environmentally exposed populations. Radiation exposure is therefore highly relevant to healthcare analytics, both in clinical risk modeling and environmental health surveillance.

Why Radiation Surveillance Data is Healthcare-Related

Environmental and occupational exposure to ionizing radiation is a major public health concern due to its well-documented role in causing cancer and other diseases. Monitoring and analyzing radiation data are critical for assessing risk, guiding policy, and protecting communities and workers.

Recent research underscores the connection between environmental radionuclides and health impacts (Egbueri et al., 2025), demonstrates molecular mechanisms and health outcomes of exposure (Shakyawar et al., 2023), and links community radon exposure to lung cancer incidence (Rosenberger et al., 2024).

```
sc <- spark_connect(master = "spark://spark-master:7077")
```

```
# Path inside Docker container (mounted from Spark host)
csv_path <- "/shared-data/radiological-air-sample-quarterly-composites.csv"
```

```
rad_tbl <- spark_read_csv(sc,
                          name = "rad_data",
                          path = csv_path,
                          header = TRUE,
                          infer_schema = TRUE)
```

```
# Check columns
colnames(rad_tbl)
```

```
## [1] "Sampling_Period"
## [2] "Sampling_Location"
## [3] "Radionuclide"
## [4] "Result_pCim3"
## [5] "Counting_Error_pCim3"
## [6] "Minimum_Detectable_Activity_95_pCim3"
## [7] "Total_Effective_Dose_Equivalent_millirem"
## [8] "NRC_Effluent_Air_Conc_Limit_pCim3"
```

```
# Clean and prepare columns
rad_tbl <- rad_tbl %>%
  mutate(
    result = as.numeric(Result_pCim3),
    dose = as.numeric(Total_Effective_Dose_Equivalent_millirem)
  ) %>%
  filter(!is.na(result), !is.na(dose))
```

Data Summary

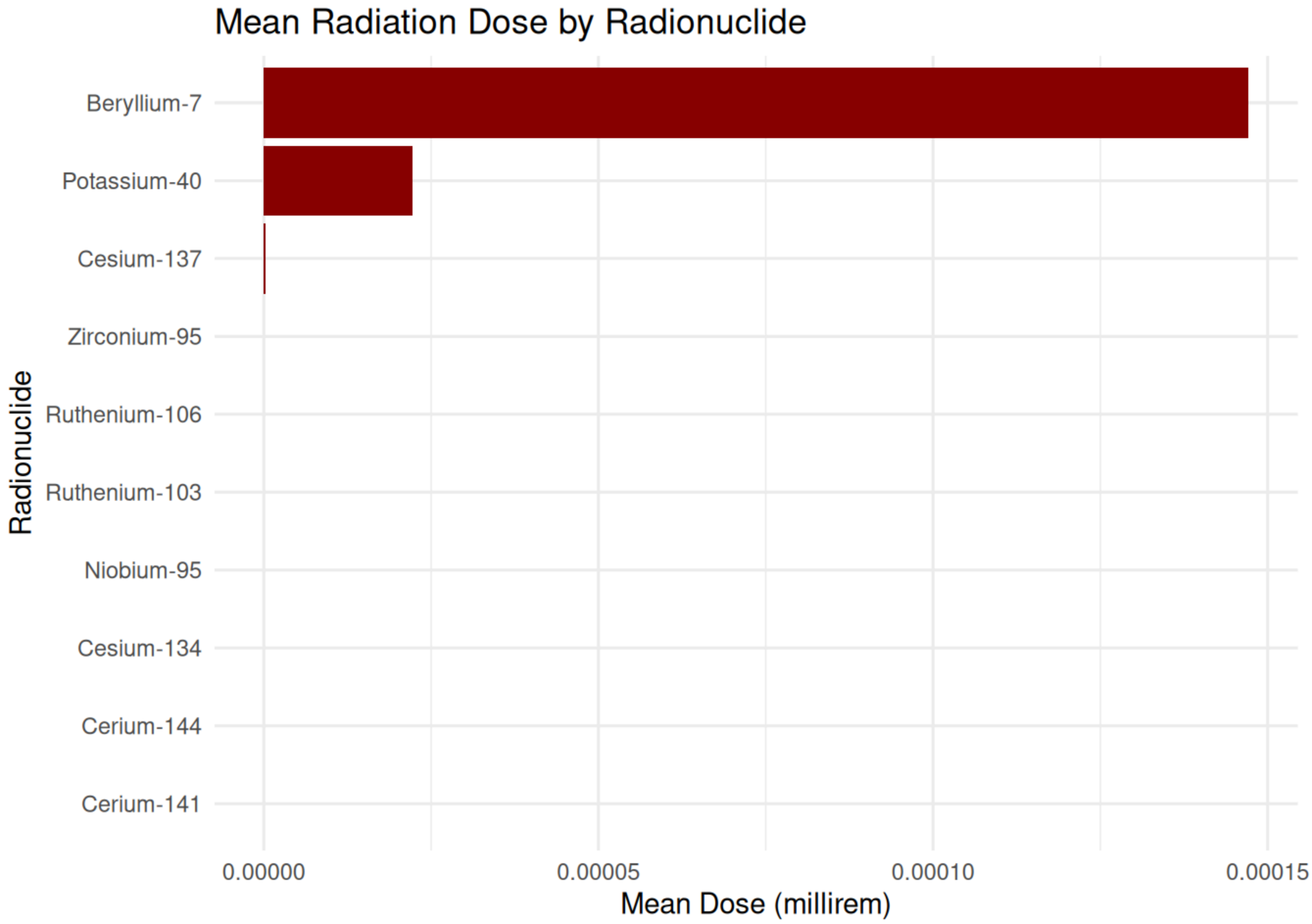
```
summary_tbl <- rad_tbl %>%
  group_by(Radionuclide) %>%
  summarise(mean_dose = mean(dose, na.rm = TRUE)) %>%
  arrange(desc(mean_dose)) %>%
  collect()

head(summary_tbl)
```

```
## # A tibble: 6 × 2
##   Radionuclide   mean_dose
##   <chr>              <dbl>
## 1 Beryllium-7  0.000147
## 2 Potassium-40 0.0000221
## 3 Cesium-137   0.000000285
## 4 Cerium-144   0
## 5 Zirconium-95 0
## 6 Cesium-134   0
```

Radiation Doses

```
ggplot(summary_tbl, aes(x = reorder(Radionuclide, mean_dose), y = mean_dose)) +
  geom_col(fill = "darkred") +
  coord_flip() +
  labs(title = "Mean Radiation Dose by Radionuclide",
       x = "Radionuclide", y = "Mean Dose (millirem)") +
  theme_minimal()
```



Feature Prediction for TEDE

```
# Feature engineering
rad_model_tbl <- rad_tbl %>%
  ft_string_indexer("Radionuclide", "rad_idx") %>%
  ft_string_indexer("Sampling_Location", "loc_idx") %>%
  ft_vector_assembler(
    input_cols = c("rad_idx", "loc_idx", "result"),
    output_col = "features_vec"
  )
```

```
# Fit regression model to predict dose
model <- ml_linear_regression(
  rad_model_tbl,
  response = "dose",
  features = "features_vec"
)
```

```
# Predict
preds <- ml_predict(model, rad_model_tbl) %>%
  select(Radionuclide, Sampling_Location, dose, prediction) %>%
  head(10) %>%
  collect()

# View results
print(preds)
```

```
## # A tibble: 10 × 4
##    Radionuclide  Sampling_Location     dose prediction
##    <chr>         <chr>                <dbl>      <dbl>
##  1 Cesium-134    Eureka            0        0.00000448
##  2 Beryllium-7   Eureka            0.000137 0.000138
##  3 Potassium-40  Eureka            0        0.00000258
##  4 Niobium-95    Eureka            0        0.00000261
##  5 Zirconium-95  Eureka            0        0.00000371
##  6 Ruthenium-103 Eureka            0        0.00000322
##  7 Ruthenium-106 Eureka            0        0.00000491
##  8 Cesium-137    Eureka            0        0.00000231
##  9 Cerium-141    Eureka            0        0.00000151
## 10 Cerium-144    Eureka            0        0.00000143
```

```
spark_disconnect(sc)
```

Summary

This analysis demonstrates Spark’s scalable power in environmental health modeling. By integrating R, Docker, and Spark, we efficiently analyzed radiation exposure data. Given the proven links between ionizing radiation and cardiovascular disease, such tools are essential for proactive healthcare analytics.

References

Egbueri, J. C., et al. (2025). Radionuclides as environmental contaminants of concern: Threats to public health through soil and groundwater. In P. Li et al. (Eds.), *Sustainable groundwater and environment: Challenges and solutions*. Springer Hydrogeology. https://doi.org/10.1007/978-3-031-82194-3_15 (https://doi.org/10.1007/978-3-031-82194-3_15)

Shakyawar, S. K., et al. (2023). A review of radiation-induced alterations of multi-omic profiles, radiation injury biomarkers, and countermeasures. *Radiation Research, 199*(1), 89–111. https://doi.org/10.1667/RADE-21-00187.1 (https://doi.org/10.1667/RADE-21-00187.1)

Rosenberger, A., et al. (2024). On the informative value of community-based indoor radon values in relation to lung cancer. *Cancer Medicine, 13*, e70126. https://doi.org/10.1002/cam4.70126 (https://doi.org/10.1002/cam4.70126)