Methodology Report: Baseline Conflation of NYSDOT RIS and OpenStreetMap Road Networks (v2 - Non-Technical)

NOTE: The following was generated using Google's Gemini (2.5 Pro Model).

Project Goal:

This project aims to spatially integrate road network information from two distinct sources: the official NYSDOT Roadway Inventory System (RIS) and the collaboratively mapped OpenStreetMap (OSM).
The primary objective of this initial phase is to establish a baseline linkage between these datasets using automated map matching.
This creates a foundational dataset where OSM road segments are annotated with corresponding NYSDOT route identifiers and attributes.
This baseline serves as a starting point for ongoing road network resilience research, allowing for iterative improvements and validation of the conflation methodology.

1. Data Sources and Preparation:

- **OpenStreetMap (OSM) Network:**
  A detailed digital road network for Albany County, NY (including a buffer area) was generated using OSM data.
  This involved extracting road geometries, building a network graph suitable for routing analysis, and topologically simplifying it.
  Simplification means that road sections between significant intersections are represented as single segments in the final network dataset used for analysis, while information about the original, more detailed OSM ways comprising that segment is retained for reference.
  These simplified segments were enhanced with attributes like length, estimated travel time, road classification, and detailed references to the original OSM data.
  The final OSM network representation consists of intersection points (nodes) and the simplified road segments connecting them (edges).
- **NYSDOT Roadway Inventory System (RIS):**
  The core RIS dataset provides geometries (linestrings) for road segments as defined by NYSDOT, referenced by a unique `ROUTE_ID` and associated with a linear referencing system (milepoints).
  A key attribute is the `DIRECTION` code, which indicates how the geometry and milepoint system relate to traffic flow (e.g., representing the primary direction, the reverse direction on divided

highways, or potentially both directions on undivided roads).
Associated RIS attribute tables detailing Functional Classification, STRAHNET status, Truck Route designations, Traffic Count statistics (like AADT), and the locations of bridges and large culverts along these routes were also utilized. These attributes are linked to specific milepoint ranges on the RIS routes.

- **NYSDOT Structures Data:**
  Detailed inventory datasets for individual bridges and large culverts maintained by NYSDOT, containing information on ownership, materials, dimensions, and condition ratings, were incorporated.

2. Map Matching Process (Linking RIS to OSM):

- **Objective:** To find the most likely path within the detailed OSM network that corresponds to the geometry of each RIS road segment.
- **Tool:** The Open Source Routing Machine (OSRM) `match` service was employed. OSRM takes a sequence of geographic coordinates (from the RIS segment geometry) and identifies the most probable sequence of connected road segments in the underlying OSM network graph that the original coordinates represent.
- **OSRM Methodology (Hidden Markov Model - HMM):** Conceptually, OSRM's matching algorithm uses a probabilistic approach, often based on Hidden Markov Models (HMMs). It considers:
  - *Observation Probability:* How likely is it that a specific input coordinate point corresponds to a particular nearby OSM road segment? This considers factors like the distance from the point to the road.
  - *Transition Probability:* Given that a point matched to OSM segment A, how likely is it that the *next* point matches to an adjacent OSM segment B? This considers network connectivity, turn restrictions, travel distance, and time between points.
  - The HMM finds the sequence of OSM road segments that maximizes the overall probability of observing the input coordinate sequence.
- **Handling Directionality:** Because the digitized direction of RIS segments doesn't always align with the primary direction of travel, the matching process was adapted based on the RIS `DIRECTION` code to improve results:
  - Segments potentially representing two-way roads (`DIRECTION=0`) were matched using both their original and reversed coordinate sequences.
  - Segments representing primary directions (`DIRECTION=1`) were matched using their original sequence.
  - Segments representing reverse directions (`DIRECTION=2` or `3`) were matched using the reversed sequence.
- **Output:** Successful matches yielded a sequence of original OSM node identifiers representing the path found by OSRM, along with a confidence score indicating the reliability of the match.

3. Associating Matches with the Simplified OSM Network:

- **Goal:** Link the OSRM match results (tied to original OSM nodes and a specific RIS `ROUTE_ID`) to the *simplified* OSM road segments used in the primary analysis dataset.
- **Baseline Linking Strategy:** For this initial phase, a simple association was made:
  - The system checked if any pair of consecutive nodes from an OSRM matched path was part of the underlying original OSM structure of a simplified edge.
  - If such a link was found, the corresponding RIS `ROUTE_ID` was associated with that simplified OSM edge.
- **Selecting the Best RIS Route per OSM Edge:** A simplified OSM edge might correspond to node pairs from multiple different RIS route matches (especially near overlaps or complex interchanges). To establish a one-to-one baseline relationship, a scoring method selected the *single best* RIS `ROUTE_ID` match for each simplified OSM edge. This score prioritized matches that covered a larger proportion of the simplified edge's length and had higher OSRM confidence scores, aiming for a score of 1.0 as the ideal. A slight penalty was applied if the match came from reversing a `DIRECTION=0` RIS segment, favoring explicitly defined reverse segments (`DIRECTION=3`) if available.
- **Handling Unmatched Directions on Two-Way Roads:** In cases where the simplified OSM network contained edges representing both directions of travel for a road (e.g., `(A, B)` and `(B, A)`), but OSRM only found a good match for one direction (e.g., `(A, B)` matched RIS Route X), the system assigned the same RIS Route X information to the unmatched direction (`(B, A)`), adjusting the associated milepoint range accordingly. This provides a more complete baseline linkage for two-way segments.

4. Integrating NYSDOT Attributes onto OSM Segments:

- **Process:** Once each relevant simplified OSM edge was linked to a single best-matching RIS `ROUTE_ID` and an estimated corresponding milepoint range on that route, attributes from the various NYSDOT RIS tables and structures datasets were joined onto the OSM edges using spatial overlap logic executed via DuckDB SQL queries.
- **Attribute Selection (Events):** For linear event data like Functional Class, STRAHNET status, Truck Routes, and Traffic Counts, the attribute value chosen for an OSM edge was taken from the NYSDOT record whose milepoint range had the **maximum overlap** with the estimated milepoint range associated with the OSM edge's match. This assumes attributes are generally constant between major intersections.
- **Attribute Selection (Structures):** For point-like structure data (bridges, large culverts) that might overlap with an OSM edge's milepoint range, the system prioritized linking the structure with the **worst (lowest) condition rating**, highlighting potential network vulnerabilities.

5. Final Conflated Output (`final_gdf`):

- The final output is a geographic dataset (`final_gdf`) where each record represents a simplified road segment from the OSM network.
- It contains the OSM segment's geometry, its unique identifiers (`u` , `v` , `key`), the linked NYSDOT RIS `ROUTE_ID` , the estimated corresponding RIS milepoint range (`min_from_mi` , `max_to_mi`), and the integrated attributes (Functional Class, AADT, Truck Route status, overlapping structure information, etc.) selected based on the maximum overlap or condition rating criteria.

Limitations and Future Directions:

This methodology provides a valuable baseline linkage but has recognized limitations inherent in automated matching and the MVP approach. Future work aims to improve accuracy and robustness by: incorporating NYSDOT calibration points to correct milepoint drift; pre-processing RIS geometries to handle discontinuities; refining the match selection heuristics; and potentially adding other matching techniques (Knowledge Sources) alongside OSRM within a more sophisticated integration framework.

Data Dictionary: `final_gdf` GeoDataFrame

This dictionary describes the columns in the final output GeoDataFrame (`final_gdf`), derived from the process in `create_ris_gdf.pdf` and referencing the provided NYSDOT documentation.

Index:

- `u` (int): Start node ID (OSMnx simplified graph) for the road segment.
- `v` (int): End node ID (OSMnx simplified graph) for the road segment.
- `key` (int): Edge key (OSMnx, usually 0 after simplification).

Columns:

- **`geometry`** (geometry): Linestring geometry of the simplified OSM road segment (CRS: EPSG:4326).
- **`_idx_`** (int): Internal row identifier from intermediate processing.
- **`route_id`** (varchar): The NYSDOT RIS `ROUTE_ID` matched to this OSM segment.
- **`min_from_mi`** (double): Estimated starting milepoint on the `route_id` corresponding to this OSM segment's matched portion.
- **`max_to_mi`** (double): Estimated ending milepoint on the `route_id` corresponding to this OSM segment's matched portion.

- **`ris_fclass_overlap_dist_mi`** (double): Length (miles) of overlap between the matched milepoint range and the selected Functional Class record's range.
- **`ris_fclass_nysdot_fclass`** (int64): NYSDOT Functional Classification code. Selected based on maximum overlap. Codes:
  - `01` : Rural - Principal Arterial - Interstate
  - `02` : Rural - Principal Arterial - Other Freeway/Expressway
  - `04` : Rural - Principal Arterial - Other
  - `06` : Rural - Minor Arterial
  - `07` : Rural - Major Collector
  - `08` : Rural - Minor Collector
  - `09` : Rural - Local
  - `11` : Urban - Principal Arterial - Interstate
  - `12` : Urban - Principal Arterial - Other Freeway/Expressway
  - `14` : Urban - Principal Arterial - Other
  - `16` : Urban - Minor Arterial
  - `17` : Urban - Major Collector (NOTE: Appears mislabeled as 'Major' in manual page 89, likely Urban Collector based on context/FHWA mapping)
  - `18` : Urban - Minor Collector (NOTE: Appears mislabeled as 'Minor' in manual page 89, likely Urban Collector based on context/FHWA mapping)
  - `19` : Urban - Local
  - `00` : Not a Highway
- **`ris_fclass_fhwa_fclass`** (int64): Derived FHWA Functional Classification code. Mapping from NYSDOT code: 1->1, 2->2, 4->3, 6->4, 7->5, 8->6, 9->7. Selected based on maximum overlap.
- **`ris_fclass_federal_aid_flag`** (varchar): Federal Aid system relationship associated with the functional class. (Details in RIS documentation).
- **`ris_fclass_class_hierarchy`** (varchar): Functional class hierarchy level. (Details in RIS documentation).
- **`ris_strahnet_overlap_dist_mi`** (double): Overlap distance (miles) with the selected STRAHNET record.
- **`ris_strahnet_code`** (varchar): Strategic Highway Network (STRAHNET) code. Selected based on maximum overlap. Codes:
  - `0` : Not a STRAHNET route
  - `1` : Interstate STRAHNET route
  - `2` : Non-Interstate STRAHNET route
  - `3` : STRAHNET Connector route
  - `N` : Not a highway
- **`ris_strahnet_description`** (varchar): Text description of the STRAHNET code.

- **`ris_trk_rte_overlap_dist_mi`** (double): Overlap distance (miles) with the selected Truck Route record.
- **`ris_trk_rte_code`** (varchar): Truck Route designation code. Selected based on maximum overlap. (Codes defined in RIS Domain Dictionary).
- **`ris_trk_rte_code_description`** (varchar): Text description of the Truck Route code.
- **`ris_bridge_overlap_dist_mi`** (double): Overlap distance (miles) with the selected Bridge Inventory record.
- **`ris_bridge_bin`** (varchar): Bridge Identification Number (BIN) of the overlapping bridge structure. Selected based on lowest condition rating.
- **`ris_nysdot_bridge_carried`** (varchar): Feature carried by the bridge. (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_crossed`** (varchar): Feature crossed by the bridge. (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_primary_owner`** (varchar): Primary owner code. (Codes: 10=NYSDOT, 30=County, 40=Town, 41=Village, 42=City, etc.). (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_primary_maintainer`** (varchar): Primary maintaining agency code. (Codes same as owner). (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_gtms_structure`** (varchar): Bridge General Type Main Span structure code (e.g., 01:Slab, 02:Stringer, 03:Girder/Floorbeam, 05:Box Beam Multiple, 06:Box Beam Single, 09:Truss-Deck, 10:Truss-Thru, 11:Arch-Deck, 19:Culvert). (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_gtms_material`** (varchar): Bridge General Type Main Span material code (e.g., 1:Concrete, 3:Steel, 5:Prestressed Concrete, 7:Timber). (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_number_of_spans`** (int): Number of spans in the bridge. (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_condition_rating`** (float): NYSDOT bridge condition rating (1.0 - 7.0 scale, lower=worse). (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_bridge_length`** (float): Total length of the bridge structure (feet). (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_deck_area_sq_ft`** (float): Bridge deck area (square feet). (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_aadt`** (int): Average Annual Daily Traffic reported for the bridge structure. (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_year_built`** (int): Year bridge was built. (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_posted_load`** (float): Posted load limit (tons) or code (88=R-Permit Restricted, 98=Closed Construction/Seasonal, 99=Closed). (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_nbi_deck_condition`** (varchar): NBI Deck Condition Code (0-9, N). (From NYSDOT Bridges dataset).

- **`ris_nysdot_bridge_nbi_substructure_condition`** (varchar): NBI Substructure Condition Code (0-9, N). (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_nbi_superstructure_condition`** (varchar): NBI Superstructure Condition Code (0-9, N). (From NYSDOT Bridges dataset).
- **`ris_nysdot_bridge_fhwa_condition`** (varchar): Overall FHWA Condition ('Good' >= 7, 'Fair' = 5 or 6, 'Poor' <= 4) based on NBI ratings. (From NYSDOT Bridges dataset).
- **`ris_large_culvert_along_mi`** (double): Milepoint location of the overlapping large culvert structure. (From RIS `Ev_Str_LargeCulvert`).
- **`ris_large_culvert_cin`** (varchar): Culvert Identification Number (CIN) of the overlapping large culvert. Selected based on lowest condition rating.
- **`ris_nysdot_large_culvert_crossed`** (varchar): Feature crossed by the large culvert. (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_primary_owner`** (varchar): Primary owner code. (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_primary_maintainer`** (varchar): Primary maintaining agency code. (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_gtms_structure`** (varchar): Structure type code (e.g., 19, 07). (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_gtms_material`** (varchar): Material type code. (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_number_of_spans`** (int): Number of spans/barrels. (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_condition_rating`** (float): NYSDOT condition rating (1.0 - 7.0 scale). (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_type_max_span`** (varchar): Code for type of maximum span. (Details likely in BDIS documentation).
- **`ris_nysdot_large_culvert_year_built`** (int): Year built. (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_abutment_type`** (varchar): Abutment type code. (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_stream_bed_material`** (varchar): Stream bed material code. (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_maintenance_responsibility_primary`** (varchar): Primary maintenance agency code. (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_abutment_height`** (float): Abutment height (feet). (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_culvert_skew`** (int): Skew angle (degrees). (From NYSDOT Large Culverts dataset).

- **`ris_nysdot_large_culvert_out_to_out_width`** (float): Out-to-out width (feet). (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_span_length`** (float): Span/barrel length (feet). (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_structure_length`** (float): Total structure length along roadway (feet). (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_general_recommendation`** (varchar): General recommendation code (1-7 scale). (From NYSDOT Large Culverts dataset).
- **`ris_nysdot_large_culvert_regional_economic_development_council`** (varchar): REDC code. (From NYSDOT Large Culverts dataset).
- **`ris_counts_overlap_dist_mi`** (double): Overlap distance (miles) with the selected Traffic Count record.
- **`ris_counts_federal_direction`** (int64): Federal direction code for count (1=N, 3=E, 5=S, 7=W, 9=N/S, 0=E/W). Selected based on maximum overlap.
- **`ris_counts_full_count`** (varchar): 'Y' if count covers full roadway width, blank if directional.
- **`ris_counts_oneway_flag`** (varchar): 'Y' if segment is one-way.
- **`ris_counts_calculation_year`** (int64): Year the AADT/DHV statistics apply to.
- **`ris_counts_aadt`** (int64): Annual Average Daily Traffic.
- **`ris_counts_dhv`** (int64): Design Hour Volume (30th highest hour).
- **`ris_counts_ddhv`** (int64): Directional Design Hour Volume.
- **`ris_counts_su_aadt`** (int64): Single Unit vehicle (F4-F7) AADT.
- **`ris_counts_cu_aadt`** (int64): Combination Unit vehicle (F8-F13) AADT.
- **`ris_counts_k_factor`** (double): Ratio of DHV to AADT (%).
- **`ris_counts_d_factor`** (double): Ratio of peak direction volume to total volume during design hour (%).
- **`ris_counts_avg_truck_percent`** (double): Avg Weekday Truck % (F4-F13).
- **`ris_counts_avg_su_percent`** (double): Avg Weekday Single Unit % (F4-F7).
- **`ris_counts_avg_cu_percent`** (double): Avg Weekday Combination Unit % (F8-F13).
- **`ris_counts_avg_motorcycle_percent`** (double): Avg Weekday Motorcycle % (F1).
- **`ris_counts_avg_car_percent`** (double): Avg Weekday Passenger Car % (F2).
- **`ris_counts_avg_light_truck_percent`** (double): Avg Weekday Light Truck % (F3).
- **`ris_counts_avg_bus_percent`** (double): Avg Weekday Bus % (F4).
- **`ris_counts_avg_weekday_f5_7`** (double): Avg Weekday Single Unit Truck % (F5-F7).
- **`ris_counts_axle_factor`** (double): Axle correction factor.
- **`ris_counts_su_peak`** (double): Peak hour SU volume as % of AADT.
- **`ris_counts_cu_peak`** (double): Peak hour CU volume as % of AADT.
- **`ris_counts_avg_k_factor`** (double): Average K Factor for the factor group.
- **`ris_counts_avg_d_factor`** (double): Average D Factor for the factor group.

- **`ris_counts_truck_aadt`** (int64): Truck AADT (F4-F13).
- **`ris_counts_morning_highest_value`** (int64): Highest hourly volume in morning.
- **`ris_counts_afternoon_highest_value`** (int64): Highest hourly volume in afternoon.
- **`ris_counts_evening_highest_value`** (int64): Highest hourly volume in evening.

Resources

1. New York State Department of Transportation (NYSDOT).
   (2020).
   NYSDOT Bridge and Large Culvert Inventory Manual.
   New York State Department of Transportation.
   https://www.dot.ny.gov/divisions/engineering/structures/repository/manuals/inventory/NYSDOT_inventory_manual_2020.pdf
2. New York State Department of Transportation (NYSDOT).
   (2018).
   NYSDOT BDIS Inventory Record Code Guidance.
   New York State Department of Transportation.
   https://www.dot.ny.gov/divisions/engineering/structures/repository/manuals/BDIS_Inventory_Record_Code_Guidance_5-30-2018.pdf
3. New York State Department of Transportation (NYSDOT), Highway Data Services Bureau.
   (n.d.).
   Count Statistics Field Definitions.
   New York State Department of Transportation.
   https://www.dot.ny.gov/divisions/engineering/technical-services/highway-data-services/hdsb/repository/Count_Statistics_Field_Definitions.pdf
4. Saki, S., & Hagen, T.
   (2022).
   A Practical Guide to an Open-Source Map-Matching Approach for Big GPS Data.
   SN Computer Science, *3*(5), 415.
   https://doi.org/10.1007/s42979-022-01340-5
5. Berkeley Institute for Data Science (DLAB).
   (n.d.).
   A brief primer on Hidden Markov Models.
   DLAB Blog.
   https://dlab.berkeley.edu/news/brief-primer-hidden-markov-models
6. Valhalla Team.
   (n.d.).

Map Matching in a Programmer's Perspective.
Valhalla Documentation.
https://valhalla.github.io/valhalla/meili/algorithms/

7. Wöltche, A.
(2023).
Open source map matching with Markov decision processes: A new method and a detailed benchmark with existing approaches.
Transactions in GIS, *27*(7), 1959–1991.
https://doi.org/10.1111/tgis.13107