ANALYSING RAILWAY INFRASTRUCTURE UTILISATION: A GEOSPATIAL APPROACH

Stuart Gordon

Submitted in partial fulfilment of the requirements for the degree of Master of Research in Geospatial Data Science in the Faculty of Science Agriculture and Engineering

Geospatial Systems Centre for Doctoral Training

School of Engineering

Newcastle University

NE1 7RU

United Kingdom

25th August 2023

Declaration:

I hereby certify that this work is my own, except where otherwise acknowledged, and that it has not been submitted previously for a degree at this, or any other university.

Stuart Gordon

25th August 2023

Acknowledgements

I would like to extend my gratitude to my supervisors, Dr Craig Robson, Professor Roberto Palacin and Oliver Bratton from Network Rail for their support and guidance.

Special mention is due to the pioneering domain of generative AI, which has significantly enhanced both the writing and code development processes. This project stands as a testament to the transformative role of AI in research.

This project is funded by Network Rail, UK, and the Engineering and Physical Sciences Research Council (EPSRC), which is a part of UK Research and Innovation (UK RI). The project contributes towards to my MRes in Geospatial Data Science, conducted as part of Studentship in the EPSRC Centre for Doctoral Training (CDT) in Geospatial Systems (ref EP/S023577/1).

Lastly, I'd like to acknowledge all the unseen contributors for their tireless efforts and unwavering support.

Abstract

In the evolving landscape of transportation, the optimisation of railway infrastructure utilisation emerges as a pivotal concern and a significant opportunity to enhance public transport. This project dives into the intricacies of railway infrastructure utilisation, focusing on the challenges of determining train service paths using both traditional methods and innovative predictive models. The study reviews methods from current literature for deducing network utilisation and capacity, with a particular emphasis on validating these findings against empirical data and use of real-world examples. Adopting a geospatial approach, the research calculates train paths to determine infrastructure utilisation and conducts a comparative analysis between the actual and planned utilisation of the train service. The results revealed a range of outputs, with discrepancies primarily due to data quality challenges. The unexpected inclusion of industrial action during the data collection period showcased the system's adaptability and provided a unique operational perspective. The comparisons between a typical and degraded service in different geographical areas underscored the methodology's potential even with data limitations. Incidents were not able to be detected though the comparison between planned and actual utilisation methodology with current data limitations however use of the probabilistic model in an alternative use could achieve this. To accommodate the expansive UK network, technologies such as NoSQL databases and serverless cloud native concepts, were employed to amplify scalability. While the project achieved most of its goals, it underscores the continual need for validation given the vast and dynamic

nature of the system and its sensitivity to data quality and availability when monitoring large dynamic real world systems.

Key Words: Railway Utilisation, Geospatial Networks, Capacity Analysis, Digital Twin, Event Detection, Cloud Computing, Railway Systems, Big Data.

Contents

List of Figures

All figures are created by the author unless otherwise stated. Figures from the Appendix are omitted from this list of figures.

List of Tables

All tables are created by the author unless otherwise stated. Tables from the Appendix are omitted from this list of tables.

1 Introduction, Aims and Objectives

1.1 The UK Railway Landscape

The United Kingdom's railway system, recognised for its rich history and global significance, forms a critical component of the nation's transportation and economic structure. Despite its importance, it grapples with considerable challenges related to performance, capacity, and resilience. A report by the Department of Transport (Department for Transport 2018) noted that in 2018/19, only 62.5% of trains arrived within one minute of their scheduled time, suggesting resilience limitations in managing and recovering from delays. Such systemic vulnerability underscores the potential for significant perturbations during events such as industrial action, infrastructure problems or changes, service limitations (including train and crew availability) or timetable complications.

In tandem, capacity constraints are an ongoing concern, exacerbated by continuous passenger growth, especially in major cities like London and Birmingham, as pointed out by the Office of Rail and Road Annual Review of 2020 (Office of Rail and Road 2020). Infrastructure demands have led to the launch of Capacity Enhancement Plans, including those for the Castlefield Corridor (Network Rail 2021a). Hence, to maintain its central role in the socio-economic fabric of the nation, the UK railway system must address these formidable challenges.

1.2 Strategies for Enhancing UK Railway Performance

In December 2022, the High Level Output Specification Control Period 7 (CP7) for 2024 to 2029 was introduced (Department for Transport 2022), prioritising evidence-based improvements to railway performance. With escalating climate hazards, the enhancement of public transportation has become increasingly critical. Proposals for augmenting rail services in the future include increasing service frequency and minimising disruptions, which could enhance user satisfaction and service usage (Transport Committee 2016).

Whilst novel infrastructural developments like High Speed 2, East Coast Digitisation Program, and Crossrail can significantly elevate performance, they necessitate substantial capital investment (*£1 billion technology investment to bring railway into 21st century* 2022; Greater London Authority 2020; HS2 2023). An alternative approach involves the strategic application of analytical methods to improve infrastructure utilisation, examining timetable implementation and overall efficiency. Though this method may not yield as substantial performance boosts as infrastructure development, it requires less capital investment and could yield gradual enhancements, particularly within the short-term control period.

This approach aligns with the UK Government's Geospatial Strategy for 2020 to 2025 (*The UK's Geospatial Strategy, 2020 to 2025* 2020), which seeks to harness location-based data to realise objectives in several areas, including railway infrastructure. This strategy identifies six core trends: real-time data, widespread sensor use, artificial intelligence, cloud and edge computing, connectivity and 5G, and data visualisation.

Network Rail's System Operator, responsible for timetable production, published a Data Architecture Reference Model (DARM) in 2019 (System Operator 2019), emphasising the crucial interdependence between timetable and railway performance and the data quality. The DARM elucidates numerous data-related issues that obstruct performance improvement.

1.3 Impact of Climate Change

In addition to operational challenges, the resilience of the UK Railway extends to managing severe weather conditions induced by climate change, which frequently disrupt services as described in Network Rail Weather Resilience and Climate Change Adaptation Report (Network Rail 2021b).

The severity of climate hazards is increasing, and efforts to reduce carbon emissions through improved public transportation resonate with the UN Sustainable Development Goals and Targets. A deeper understanding of existing rail infrastructure and services utilisation can minimise the need for large-scale construction projects, thereby enhancing the appeal of rail services as a sustainable public transportation option. Such advancements could notably contribute to reduced carbon emissions and improved public transportation choices.

1.4 Project Aims & Objectives

In alignment with the aforementioned policy concepts, this project intends to scrutinise and assess the utilisation of the UK Railway, using the trends identified in the geospatial strategy to address the highlighted data issues.

The project aims to:

- Examine the utilisation of railway infrastructure, contrasting the planned path and the actual path of train services.
- Analyse differences between the planned and actual timetables of trains.
- Address the data deficiencies in determining the train's path
- Determine a mechanism for monitoring current infrastructure utilisation.
- Detect incidents through the analysis of the proposed method's results, both retrospectively and in real-time.

The project objectives include:

- Constructing a geospatial infrastructure by amalgamating geographical and logical network data.
- Calculating the geospatial path of each operating train service and forecasting the time-based occupations of the network.
- Assessing the utilisation of geospatial infrastructure based on the projected paths of the train services.

- Comparing the planned and actual timetable and path of each train service and evaluating the resultant differences in utilisation.

- Verifying the proposed system against real-time train operations to identify any model misalignment's.

2 Background

The railway system in the UK has evolved significantly, marked by a rich tapestry of technological advancements, operational challenges, and consumer demands. This background section aims to provide readers with a comprehensive understanding of various facets of the UK railway system, specifically examining train timetables, rail utilisation, and the intricate data layers that underpin the system's operations. Initial discussions will delve into the broader realm of train timetables and how the UK's railways are utilised. Subsequently, an in-depth exploration of the various data layers of the UK railway system will be presented, from timetables and planning geography to sophisticated components like railway signalling control systems and the integration of Train Running Under System TOPS (TRUST), the networks system for monitoring operational performance and delays. Furthermore, a focused examination of these data layers, centred around the iconic Welwyn Viaduct, will offer insights into the railway systems real-world operations and complexities .

2.1 Train Timetables

Within the UK railway, the production of train timetables is the responsibility of the System Operator within Network Rail. These timetables specify the journey times from one location to another at a 30-second resolution (Network Rail 2017). These intervals between locations represent the agreed average time for that section, known as the Sectional Running Time (SRT). The development of these timetables is governed by the Timetable Planning Rules. The timetable is loosely coupled to the train movements — implicitly stating the path each train will take — at the specification stage, which is approximately 55 weeks prior to timetable implementation. This modelling primarily focuses on validation of the specification, generally not representing an actual timetable due to adjustments to be made as late as two days before the planned implementation which are not modelled. A summary of this process and the associated process maps are defined in the Data Architecture Reference Model (DARM) (System Operator 2019).

These changes are generally not planned to the train movement level—the path each train will take—which can further compound the issue of conflicts with other services leaving their actual path taken up to the discretion of the signaller, responsible for the safe movement of trains in their area of control.

2.2 UK Rail Utilisation

In the UK's railway network, train movements are inherently governed by the timetable, while the explicit path between signals is largely under the signaller's control. This design allows a degree of adaptability, permitting signallers to alter train routing in response to daily operational fluctuations. A signaller is an operator responsible for the safe movement of trains on the network by controlling the infrastructure through a control system.

Historically, this approach was sufficient as human signallers could implicitly infer timetables. This remained largely unchanged from the 1900s through to the early 2000s. However, due to increasing demands stemming from elevated train frequency and expanded control areas for signallers due improvements in technology. Automated solutions have emerged to manage this escalating workload. These systems have been the subject of numerous Human Factors and Ergonomics investigations in recent years (Balfe et al. 2011; Krehl and Balfe 2014; Pickup, Wilson, and Lowe 2010). These systems deduce a trains route from the timetable, guiding trains as intended with best effort. Due to the need for the source data to be significantly cleansed and corrected, these systems are typically limited to small regions due to the labour-intensive nature of data production. The ORR detailed examples of such issues related to the adoption of technology in 2022 (Office of Rail and Road 2014). National implementations are envisaged to be impracticable due to the substantial data requirements, lack of high quality data and the obstacles in addressing these data limitations.

The current flexibility in the timetabling strategy, however, suggests that optimal infrastructure usage is neither preplanned nor consistently achieved during daily operations. It varies based on the signaller's current competence and is reflective of current human performance (Balfe et al. 2011). Despite a handful of nascent automated systems beginning to explore the application of optimisation algorithms, very few have been implemented, focus on delay reduction, and those that exist are based on 1970s technology concepts.

A notable absence in the UK railway system is a utilisation performance metric, which would help measure how efficiently the infrastructure is used. The resolution of such a metric could vary greatly, ranging from a specific track section to an entire line of route, which could significantly influence the calculation of utilisation, similar to the concept of the Modifiable Areal Unit Problem (MAUP) (Wong 2009). One reason for this lack of a utilisation metric may align with the issues outlined in the DARM summary (System Operator 2019). A key issue with improving the timetable and the ability to systematically process data (and calculate metrics) is the lack of an industry data platform or standard data practices.

2.3 Data Layers of UK Railway

A significant number of systems are required to track and monitor the UK railway network. These systems can be imagined to exist as layers with their own digital assets and data in each layer. These layers are typically hierarchical but without linkages between the layers. This is described as one of the key themes in the DARM Summary (System Operator 2019). Improving these layers and the connections between these are key to the DARM theme of "Better Data, Better Timetable".

To aid in the understanding of these layers, the following subsections explain these layers followed by a worked example detailing the existing linkages and interactions between the layers.

2.3.1 Timetable

The railway timetable in the United Kingdom serves as an exhaustive schedule for trains, addressing the requirements of passengers, operational users, and systems alike. Its design and efficacy are anchored by the timetable planning rules (Network Rail 2023b), an example is shown in Figure 2.2. These rules prescribe guidelines for adherence to specific parameters such as times and locations, to produce a precise timetable.

This approach to data governance is pivotal to the timetable's effectiveness and accuracy, enhancing the overall reliability and efficiency of the railway network. This system reduces potential scheduling anomalies and operational inconsistencies, thereby streamlining the day-to-day operations of the railway network.

The creation of this daily timetable is the responsibility of Network Rail and its Train Planning System (TPS). Network Rail begins the timetable preparation process 18 months in advance, yielding a base timetable with a six-month validity, this process is defined in the DARM (System Operator 2019). This framework allows for modifications of a days timetable up to the day before its enactment, which ensures the timetable remains responsive to any unforeseen changes in daily operations. This meticulous planning and updating process underscores the timetable's crucial role in maintaining the operational efficiency of the UK railway network.

An example of the timetable is shown in Table 2.2.

2.3.2 Planning Geography

The planning geography serves as an intricate and comprehensive representation of the actual railway network. The model is maintained and disseminated by Network Rail through the BPlan system. Regarded as the "master geography", this data forms the foundational basis for train planning systems, encapsulating a vast array of information such as location identifiers, platforms, running lines, and activities. This may differ from the physical infrastructure on the network.

A geospatial model of the physical track layout is available from Network Rail, though it does not have logical features that align to the planning geography presented in BPlan, an example of this is shown in Figure 2.4.

2.3.3 Railway Signalling Control Systems

A railway signalling control system is an intricate technology infrastructure that orchestrates train movement, ensuring optimal safety and efficiency throughout the rail network. This complex system comprises an interconnection of devices, processes, and protocols. It ranges from trackside equipment such as signals and switches, and detection mechanisms, to more centralised components like computer-based interlocking and automatic train control systems. The primary objective of this system is to prevent collisions, regulate safe speeds, and facilitate efficient utilisation of tracks and stations.

In their infancy, these systems were fundamentally mechanical, controlling semaphore signals via physical means. The subsequent phase in their evolution was the transition to electro-mechanical systems, characterised by heavy reliance on relays and large physical display panels. The modern era has ushered in computer-based control systems, which leverage visual display units for enhanced control and efficiency. The layout of the visual display unit is carefully controlled by Network Rail and its defined ergonomics standards.

2.3.4 Train Describers (TD)

These systems represent a critical element within railway signalling control systems, their principal function being to track and communicate the unique identity, or 'headcode', of a train as it navigates the signalling infrastructure. These TD systems are deeply integrated into the broader signalling system, not only deriving their operational logic from the state of the signalling system but also reciprocating valuable data to centralised monitoring systems. Importantly, TDs facilitate the transfer of train identity information at signalling boundaries, enabling transmissions of headcodes to neighbouring signalling systems and associated TDs, an example of this is shown in Figure 2.7. This interlinked architecture affords an effective visual depiction of trains within designated signalling areas, thus amplifying the level of situational awareness available to signallers. TDs also report event changes of the system to SMART (Signal Monitoring and Reporting to TRUST), which collects data nationally from all TD systems.

Historically, TD systems were predominantly electronic, utilising EPROM (Erasable Programmable Read-Only Memory) processors to present train identities on substantial physical signalling panels. However, the digital revolution has precipitated a profound transformation of these systems. Contemporary TDs operate via custom software hosted on standard computer-based platforms.

2.3.5 Train Running Under System TOPS (TRUST)

TRUST is a pivotal system within the UK railway infrastructure, responsible for tracking train progress and recording running information for both operational scrutiny and performance evaluation. TRUST interfaces with signalling systems, capturing real-time data related to train movements across the extensive network. The system records various train events including timetable changes, departures, arrivals, and passing times at timetabled locations.

A comparative analysis of these empirical data points against the planned train timetable enables TRUST to monitor a train's progress and report on punctuality and overall adherence to the schedule. This process is documented in the Delay Attribution Review by the Rail Delivery Group (Railway Delivery Group 2023).

Established in the 1980s, TRUST is based on an IBM 370 mainframe computers installed at Marylebone in London, with IBM command line assembler macros used for human machine interface interaction. However, TRUST has a significant limitation: a hard cap of 150 locations per train service. Where services have more than 150 locations the TRUST mainframe becomes unstable and crashes. This limitation has influenced the design of the timetable, particularly for the longest long-distance services, CrossCountry Aberdeen to Penzance services. The implicit nature of the timetable was established to mitigate this and handle the limitation by reducing the required number of locations in a timetable due to the majority of human signallers and computational limistation at the time of its inception in the mid 1970s.

2.4 Data Layers of the UK Railway: An Examination of the Welwyn Viaduct

The complex interactions and relationships among the various data layers of the UK railway system can be elucidated through the example of the Welwyn Viaduct. Situated on the East Coast Main Line (ECML), this viaduct is a railway bridge traversing the River Mimram and positioned between Welwyn Garden City and Digswell (see Figure 2.1 for a view of the viaduct). Notably, it represents the only two-track section of the ECML (typically at least 4 lines for mixed traffic use) located south of Peterborough, thereby constraining train flow to a maximum of 17 trains per hour (tph) Northbound and 19 tph Southbound as of 2020 (Network Rail 2014). Including intermittent freight paths, the viaduct accommodates approximately 700 trains per day. Owing to the intense traffic, the viaduct is widely recognised as a congestion point, often operating beyond its capacity (Network Rail 2014).

All figures and tables presented herein are derived from the inspection and analysis of real-time data personally collected by the author and supplementary information sourced from OpenTrainTimes, RealTime Trains and Charl Woodhouse Live Rail. These websites, leverage the same Network Rail Open Data sources that underpin the present project are used herein.

Figure 2.1: An aerial image of the Welwyn Viaduct showing a train crossing the restricted 2 track railway viaduct. Image source: BG Droneshots

2.4.1 Example Timetable

Our example will revolve around service 1P93 (X02209 on 21/07/2023), a Great Northern service journeying from Peterborough to Kings Cross. 1P93 is the trains headcode, X02209 is the trains timetable UID, neither identity is unique on any day but in combination of both and the day of running produce a unique identity. A representation of the service's timetable can be found in Table 2.1. This user-centric representation provides a simplified view, showcasing arrival and departure times at the designated stop locations, with time measurements given at a resolution of 30 seconds.

Additionally, the table offers a glimpse into the train's route. Here, 'Platform' signifies the station platform, 'Path' reflects the route from which the train has arrived, while 'Line' represents the route that the train will follow. The specific line is not explicitly defined, though one can infer from this example (Table 2.1) that they may refer to the 'Fast Line' (FL) or the 'Slow Line' (SL) and thereby identify the explicit tracks that the train is planned to use.

However, Table 2.1 fails to include several pass-through or junction locations that appear in the fully published timetable. Of the 19 locations listed in the original timetable, only 6 serve as actual stops for the train. For a more comprehensive view, including passing locations (marked as departures without corresponding arrivals), please refer to Table 2.2.

Upon examining the train planning rules for this area (Figure 2.2), as well as the geographic layout depicted in Figures 2.4 and 2.3, we note that Knebworth is absent from the timetable. This is a result of the location being Non-Mandatory (not bold) in the TPR (Figure 2.2), in practice, the train would traverse this location at a time between 08:51:00 (departure from Stevenage) and 08:56:30 (passing Woolmer Green Junction).

Table 2.1: An example of a simplified timetable for the Great Northern 1P93 service from Peterborough to Kings Cross service, where only the stopping locations, relevant to passengers are shown. This example is from Train UID X00103 on 21/07/2023.

Location	Arr	Dep	Platform	Path	Line
Peterborough		08:05:00	2		FL
Huntingdon	08:18:30	08:19:30	2	FL	SL
St Neots	08:26:30	08:27:30	4		
Biggleswade	08:37:00	08:38:00	1		
Stevenage	08:50:00	08:51:00	1		
Kings Cross	09:17:00		9		

LN101 LONDON KING'S CROSS TO SHAFTHOLME JUNCTION				
TIMING POINT	**DOWN**	**UP**	**CODE**	**NOTES**
Welham Green	SL	SL	S	
Marshmoor		FL	X	
Hatfield	FL SL	SL FL	S X	Trains routed UWF from Welwyn Garden City via Flyover are to be timed at Hatfield
Welwyn GC Rev Sidings	–		S	
Welwyn GC Signal K167	–		S	
Welwyn GC Signal K168		–	S	
Welwyn Garden City	FL SL	FL SL UWF		Platform details must be shown for stopping trains and non–stop via platforms
Welwyn FD	FL SL	SL	S	
Welwyn Garden City Signal K180		–	S	TIPLOC (WLWY180)
Welwyn Garden City Signal K182		–	S	TIPLOC (WLWY182)
Welwyn Garden City Signal K184		–	S	TIPLOC (WLWY184)
Welwyn Garden City Carriage Sidings		–	S	
Digswell	FL	FL SL		
Welwyn North	FL	FL	S	
Woolmer Green Junction	FL SL	FL		
Knebworth	FL SL	FL SL	S	
Langley Junction	SL DL	UL DL	X	*To/from Hertford North – LN120* All trains running to/from Stevenage Platform 5 must be timed at Langley Jn

Figure 2.2: The train planning rules for the area around the Welwyn Viaduct from LNE TPR v4.3 with the mandatory and non-mandatory timetables shown in bold (Network Rail 2023a).

Table 2.2: An example timetable for the Great Northern 1P93 service from Peterborough to Kings Cross service, this example is from Train UID X02209 on 21/07/2023. This table has a number of fields from the timetable published on the Open Data feed omitted for readability

Location	Arr	Dep	Platform	Path	Line
Peterborough		08:05:00	2		FL
Fletton Junction		08:07:00			
Connington South Junction		08:12:00			
Huntingdon	08:18:30	08:19:30	2	FL	SL
St Neots	08:26:30	08:27:30	4		
Sandy		08:34:00	1		SL
Biggleswade	08:37:00	08:38:00	1		
Hitchin		08:46:00	1		SL
Stevenage	08:50:00	08:51:00	1		
Woolmer Green Junction		08:56:30			FL
Digswell Junction		08:58:00			
Welwyn Garden City		08:58:30			
Potters Bar		09:04:00	2		
Alexandra Palace		09:09:00			FL
Finsbury Park		09:11:30	4		
Holloway South Junction		09:13:00			SL
Kings Cross Belle Isle		09:15:00			EX
Kings Cross	09:17:00		9		

2.4.2 Planning Geography

The infrastructure's geography can be represented through various projections. Figure 2.4 provides a scaled, geographical depiction of the area of interest, showcasing the railway infrastructure's layout.

Conversely, Figure 2.3 presents a diagrammatic view of the infrastructure, akin to what a signaller would see. This visualisation is not drawn to scale, mirroring the display on signaller visual display units. It is designed to facilitate the railway's safe operation by giving prominence to areas with dense information, such as stations, over plain lines of track.

Signals, illustrated as red points, are geolocated against the track layout in Figure 2.4, echoing the representation in Figure 2.3. It should be noted that not all signals appear in the geospatial view due to data issues, which are further elaborated in 4.1.

In Figure 2.3, signals are depicted as red circles with a stick indicating the direction of operation, accompanied by their asset id or signal name (e.g. K603). Platforms are also displayed in relation to the track layout as orange areas. Further assets or information, such as train detection areas (e.g. track circuits) or points, and their current state would be visible on a signaller's display, but omitted for clarity.

The logical signal name, defined in the train describer system, is highlighted in a black square with white text in Figure 2.3. These markers, also known as the berth or signal berth, are included to delineate the relationship between elements.

On a signaller's display, these logical signal identities would remain blank unless populated by a train identity or headcode, indicating an approaching train to that particular signal.

The NIM infrastructure model, the underlying track model in Figure 2.4 are shown as the lines/edges connecting the tracks together. Each edge between nodes (tracks) are shown in a unique colour in Figure 2.4. These are connected edges allowing for a geospatial network model. The NIM model has the linear referencing system used the the railway embedded as meta data in the infrastructure model.

2.4.3 Train Paths

The path a train takes through the network can be inferred in advance, aiding signallers in setting the correct routes. A route is defined as a path between two signals, which a signaller secures to safely guide a train. As a train passes between two signals, a 'CA' event is generated in the train describer system, moving the train ID from one signal to the next to inform the signaller that the train is approaching the subsequent signal. These events are referred to as berth steps ('CA' Event Type in Table 2.3) and can be represented as a network graph, with signal berths as nodes and berth steps as edges. Additional events exist, such as interposes and cancellations (CC and CB Event Types), add or remove train IDs to the system at nodes in the network.

Focusing solely on berth steps, Figure 2.3 illustrates a network of possible berth steps. Signal berths are displayed similarly to Figure 2.6. The directionality of the steps is represented in green (from left to right - Down/Away from London) and blue (from right to left - Up/Towards London), aligning with railway signalling nomenclature: blue for 'Up' direction (towards London) symbolising sky, and green for 'Down' direction (away from London) indicating ground.

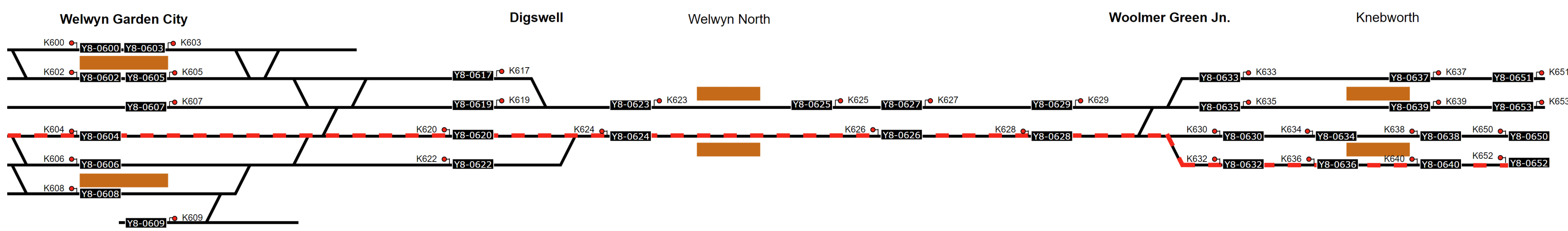


Figure 2.3: An example diagrammatic view of the signalling layout around Welwyn Viaduct from Welwyn Garden City to Knebworth. Signals are shown as circles on sticks in the direction of operation with the signal asset identifier and logical identifier (Signal Berth, white text in black box) for each signal in the area. Platforms are shown in orange with the location placed above. Only main signals are shown. The path of 1P93 taken is shown in the red dashed line over the infrastructure.

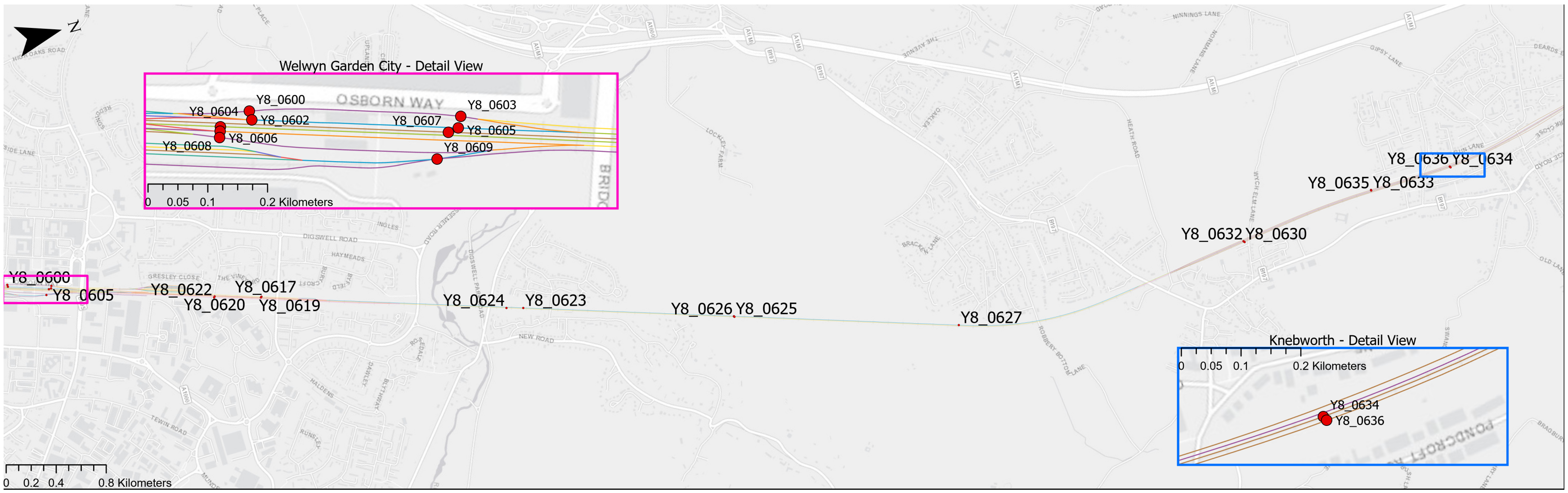


Figure 2.4: A geospatial view of the Welwyn Viaduct from Welwyn Garden City to Knebworth, with signals (red dots) geolocated onto the NIM layout in Figure 2.4 and displayed in a geographic projection. The label added to the signal is the logical id or Signal Berth as shown in Figure 2.3. The geolocated signal position data was generated using the process outlined in the 4.1.3 . Individual elements of the NIM tracks (edges of the track model) are shown in unique values, these are easiest to see in the Welwyn Garden City detailed view.

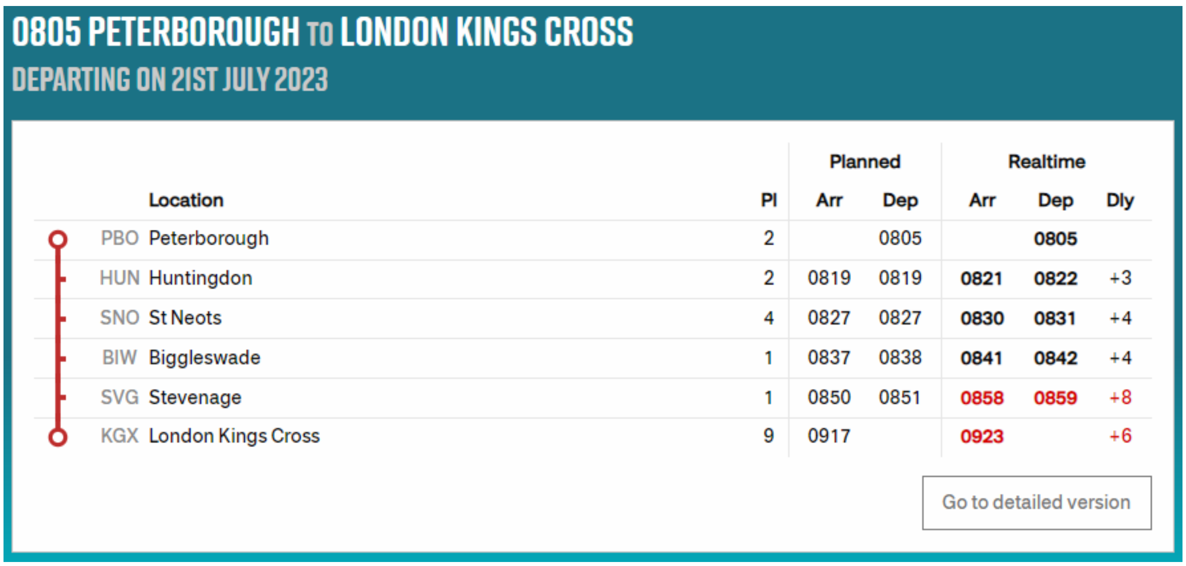
Additional colours in Figure 2.6 signify other information. Timetable locations and their associated signal berth (used by TRUST and SMART to associate train movements with its planned timetable) are grouped in red. Mandatory locations are highlighted in bold, as displayed in the TPRs (Figure 2.2). Connections between these locations, aligned with the TPRs, are represented by pink lines, indicating possible paths.

Given that Welwyn North is non-mandatory in the TPRs (Figure 2.2), the list of locations in the timetable could move directly from Digswell to Woolmer Green Jn without including Welwyn North, particularly if the train isn't stopping there. Consequently, another network can be inferred to denote the timetable under planning rules, using timetable locations as nodes and connections between locations as per the TPRs as edges.

The train's path (1P93) can be inferred from Table 2.2 and Figure 2.6. It is presumed to enter Figure 2.6 at 'Y8-0652', moving onto the fast line at Woolmer Green and continuing on the fast line at Digswell Junction ('Y8-0624' to 'Y8-0620'). However, a signaller or automated system could alter this path or order of route being set on the day based on operational needs, such as giving priority to high-speed services. The routing information provided in Table 2.2 for SL and FL are notations and is information not embedded in other layers such as TD network or NIM (or its metadata). The actual path taken by the train can be confirmed by reviewing the berth steps in Table 2.3.

Several auxiliary TD events (CC and CB Events) occur around Welwyn Garden City in Table 2.3. These events can vary due to the TD system's age, differences between TD system suppliers, and local operating practices, which may differ between locations. Typically, these events are disregarded for these analyses as they are used for platform operations of terminating services in this location which is not relevant to the example service.

Upon inspecting the step times and planned times, it is evident that 1P93 was running approximately 7 minutes late on the day of operation compared to its planned timetable which aligns to its reports on the day as seen in Figure 2.5.



0805 PETERBOROUGH TO LONDON KINGS CROSS
DEPARTING ON 21ST JULY 2023

	Location	Pl	Planned Arr	Planned Dep	Realtime Arr	Realtime Dep	Dly
PBO	Peterborough	2		0805		0805	
HUN	Huntingdon	2	0819	0819	0821	0822	+3
SNO	St Neots	4	0827	0827	0830	0831	+4
BIW	Biggleswade	1	0837	0838	0841	0842	+4
SVG	Stevenage	1	0850	0851	0858	0859	+8
KGX	London Kings Cross	9	0917		0923		+6

Go to detailed version

Figure 2.5: A summary of 1P93's timetable and actual real time performance as recorded by Real Time Trains

2.4.4 Transitions between TD Systems

As highlighted in section 2.3.4, Train Describer (TD) systems can transfer train IDs at the boundary of different signalling systems. Figure 2.7 illustrates an example of such a transition from the Kings Cross North (represented by Y8 Train Describer and Green Signal Berths) to Peterborough (represented by PB Train Describer and Blue Signal Berths). This illustration shows that signals can be given multiple berth IDs, each unique within their respective system. This aspect may generate additional

Table 2.3: An example list of TD Events for the Great Northern 1P93 service from Peterborough to Kings Cross service, this example is from Train UID X02209 on 21/07/2023. For berth steps (Type CA) the time of the step is recorded between berths and if the step corresponds to a location as would be recorded against the timetable is shown in the planned location and time columns, an offset is also added to account for duration between the step and physically passing the location which may not occur in the same spatial location.

Event Type	Event Time	From Berth	To Berth	Planned Location	Planned Time
CA	09:00:10	Y8-0658	Y8-0652		
CA	09:00:51	Y8-0652	Y8-0640		
CA	09:01:24	Y8-0640	Y8-0636		
CA	09:02:04	Y8-0636	Y8-0632		
CA	09:02:53	Y8-0632	Y8-0628	Woolmer Green Jnc.	08:56:30
CA	09:03:43	Y8-0628	Y8-0626		
CA	09:04:16	Y8-0626	Y8-0624		
CA	09:04:40	Y8-0624	Y8-0620	Digswell Jnc.	08:58:00
CC	09:04:41		Y8-WUFA		
CA	09:05:14	Y8-0620	Y8-0604		
CC	09:05:15		Y8-L604		
CB	09:05:15	Y8-WUFA			
CA	09:05:41	Y8-0604	Y8-0592	Welwyn Garden City	08:58:30
CC	09:05:41		Y8-LUFL		
CB	09:05:41	Y8-L604			
CA	09:06:30	Y8-0592	Y8-0586		
CA	09:07:02	Y8-0586	Y8-0580		

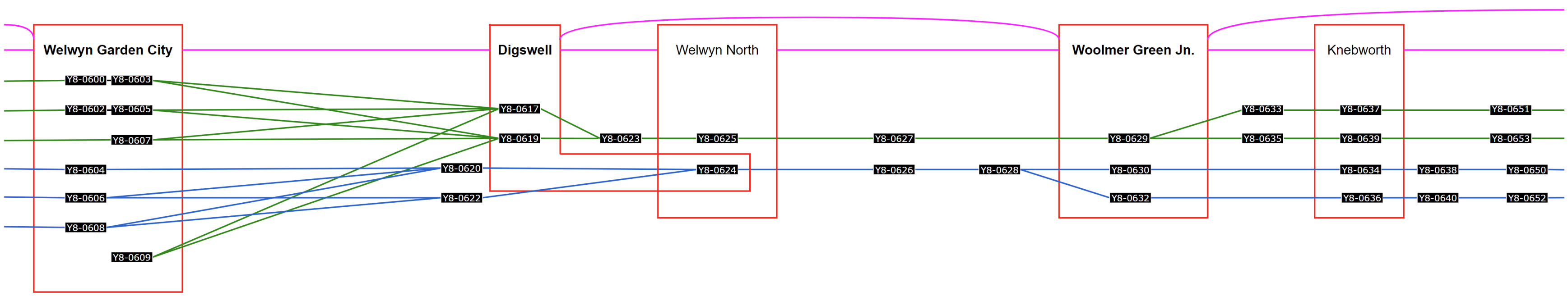


Figure 2.6: A network graph view of the Welwyn Viaduct from Welwyn Garden City to Knebworth, routes between signals are shown as edges (green lines (Down Direction) or blue lines (Up Direction)) and signal berths (black boxes) as nodes. Locations and grouping of the associated berths to that location are shown in orange borders. Edges between locations in a timetable are shown as pink lines with locations as nodes in the network graph

TD events, which can be overlooked when analysing a train's TD steps. The system behaviour and published events of these transitions varies significantly between different systems. For instance, the transitions from Y8 to PB in Figure 2.7 differ in each direction. Notably, there are no equivalents to PB-COUT at Y8-0776 & Y8-0778, yet duplicate berths are created in Y8 (Y8-PDSA for Y8-0763 & Y8-PDFA for Y8-0765). When transitioning from PB to Y8, berths are created in Y8 for PB berths for trains approaching Y8 (Y8-HUFA for PB-0276 & Y8-HUSA for PB-0274). This serves as just one example but many exist and differ across the entire network.

In practical terms, these transitions can be analysed by reviewing stepping data. For this case, the example train 1P93 is utilised, as shown in Table 2.4.

The concept of transition berths is proposed, whereby the link between two train describers can be identified. For the case of 1P93 transitioning from PB to Y8, as detailed in Table 2.4, this link would occur between Y8-0798 to Y8-0778 and PB-K798 to PB-COUT. As observed in Table 2.4, these events occur 15 seconds apart.

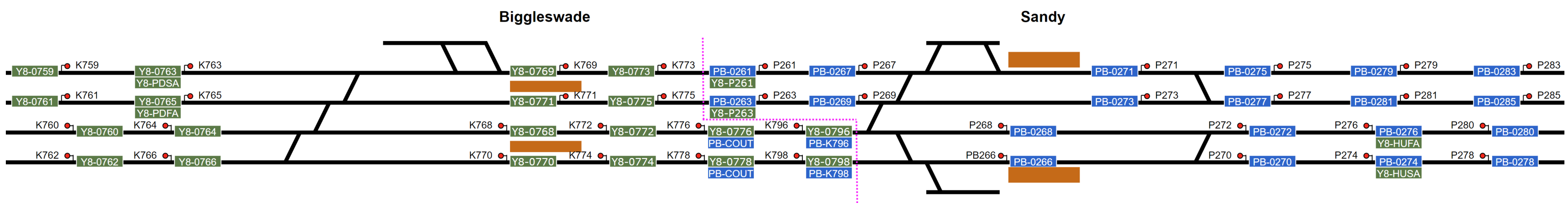


Figure 2.7: An example diagrammatic view of the signalling layout around the Kings Cross North (Y8 Train Describer - Green Signal Berths) to Peterborough (PB Train Describer - Blue Signal Berths) in the similar style to Figure 2.3 except for signal berths to clarify the different TD systems. Where an overlap or duplicate berth exists for a signal the berth has been added underneath the inferred main signal. A pink dashed line has been added to aid the theoretical boundary of the TD systems.

Table 2.4: An example list of TD transition Events for the Great Northern 1P93 service from Peterborough to Kings Cross service, from the Peterborough (PB) Train Describer to the Kings Cross North (Y8) Train Describer. This example is from Train UID X02209 on 21/07/2023.

Step Time	msg-type	From Berth	To Berth	Planned Location	Planned Time
08:35:16	CA	PB-0282	PB-0278		
08:36:01	CA	PB-0278	PB-0274		
08:36:24	CC		Y8-HUSA		
08:36:49	CA	PB-0274	PB-0270		
08:37:21	CA	PB-0270	PB-0266	Sandy	08:34:00
08:38:18	CC		Y8-0798		
08:38:19	CA	PB-0266	PB-K798		
08:38:19	CB	Y8-HUSA			
08:39:05	CA	Y8-0798	Y8-0778		
08:39:15	CA	PB-K798	PB-COUT		
08:39:37	CA	Y8-0778	Y8-0774		
08:40:54	CA	Y8-0774	Y8-0770		
08:42:49	CA	Y8-0770	Y8-0766	Biggleswade (arr)	08:37:00
08:43:42	CA	Y8-0766	Y8-0762	Biggleswade (dep)	08:38:00
08:44:21	CA	Y8-0762	Y8-0750		
08:45:05	CA	Y8-0750	Y8-0746		

3 Literature Review

This literature review aims to offer a comprehensive overview of two pivotal aspects underpinning the successful deduction and optimisation of railway infrastructure utilisation for the planned and actual path of train services: path prediction through a network or graph and methods for determining infrastructure utilisation.

A profound understanding of these facets, with insights gleaned from previous studies, provides a robust foundation for this study. Such understanding offers the capacity to address inherent challenges in the UK railway landscape, particularly those relating to the effective utilisation of infrastructure and precise path prediction.

This review firstly delves into path prediction through a graph network, addressing its fundamentals, various prediction techniques, real-world applications, and the inherent limitations of these approaches. Subsequently, the review shifts its focus towards the concept of railway infrastructure utilisation, highlighting its importance, key performance indicators, prevalent methodologies, relevant case studies, and the challenges encountered.

3.1 Predicting the Path Through a Graph Network

Navigating complex networks has become a vital challenge in various fields, with the need to predict paths through a network being increasingly significant. Networks, represented as graphs of nodes and edges, have traditionally been navigated using deterministic graph-based methods like Breadth-first Search (BFS), Depth-first Search

(DFS), or Dijkstra's Algorithm (Dijkstra 1959). However, the evolving complexity of dynamic real-world networks, where edge probabilities need estimation from uncertain data, has catalysed a shift towards machine learning and deep learning techniques for path prediction over traditional methods. This literature aims to explore, evaluate, and compare these traditional and modern approaches, identifying their potential benefits, drawbacks, and suitable use-cases, and highlighting future research directions and challenges in this rapidly advancing field.

3.1.1 Fundamentals of Graph Theory in Railways

From the background provided in section 2 a number of network graphs are layered together to describe trains paths through the infrastructure. Predicting this passage through the networks with a prescribed path for each train based on the timetable. An accurate prediction of the path is critical for ensuring efficient, timely, and safe movements. Railway networks, composed of locations (ie stations) (nodes) and tracks (edges), can greatly benefit from robust path prediction models to optimise routing and scheduling. The importance of path prediction is further emphasised by the highly interconnected and complex nature of railway networks, where delays or disruptions can have a cascading impact as described by (Cacchiani et al. 2014).

3.1.2 Path Prediction Techniques

Traditional path prediction methods for railways primarily employ graph-based algorithms such as Breadth-first Search (BFS), Depth-first Search (DFS), and Dijkstra's Algorithm. Dijkstra's algorithm, which finds the shortest path between nodes in a graph, has been extensively used in railway settings such as the Dutch railway system to determine the shortest time or distance path for a train through a network in Nachtigall 1995.

However, the deterministic nature of these traditional methods can fall short in addressing the dynamic and stochastic elements inherent to railway networks. Real-world railway systems, characterised by uncertainties like delays, track maintenance, unexpected breakdowns, and varying passenger loads, demand models that can adapt to such dynamic changes as explored by Lusby et al. 2011.

3.1.3 Predictive Models

Machine learning and deep learning techniques have shown promise in dynamic and changing environments. Models like Random Forests, Support Vector Machines (SVM), or Artificial Neural Networks (NNs) can be trained on historical data to predict future paths, evidenced in many machine learning applications.

More recently, Graph Neural Networks (GNNs), a subset of deep learning specifically designed for graph data, have been employed for forecasting paths in traffic networks (Yu, Yin, and Zhu 2018). GNNs can encode nodes, edges, and their features into continuous vectors, enabling the model to infer patterns and relationships from complex network structures.

Transformers, originally introduced in the context of natural language processing by Vaswani et al. Vaswani et al. 2017, have proven their mettle across various domains, including graph data. The self-attention mechanism of transformers allows for capturing dependencies irrespective of the distance between data points in the sequence, making them ideal for tasks where inter-node relationships are essential. In the railway context, this means understanding the relationships and inter dependencies between different stations or tracks, even if they aren't directly connected could be exploited.

3.1.4 Limitations and Future Directions

In summary, while traditional deterministic methods have long been the cornerstone for railway path prediction, the dynamic and uncertain nature of railway networks is driving a shift towards stochastic prediction models powered by machine learning and deep learning techniques.

A promising application of transformers in this field is the use of next-token prediction. Imagine representing a train's journey as a sequence of station tokens, or steps between signals. Given a partial sequence, the transformer can predict the next station or TD step the train will traverse, even down to track circuit resolution. This prediction mechanism could be used for both routing trains in real-time and forecasting future traffic on specific paths. It's worth noting that with the sequential nature of train movements, next-token prediction could be especially powerful as it takes into account both historical patterns and real-time network conditions, though this is an emerging field which has not yet been seen in the railway or application of path prediction.

As research in this field grows, plausible approach may that a hybrid, combining the strengths of both traditional and modern methods, may be the most effective strategy for future railway path prediction.

3.2 Methods for determining Railway Infrastructure Utilisation

3.2.1 Definition and Importance

Understanding and optimising infrastructure utilisation is at the heart of efficient railway operations. Before we delve deeper, it is important to distinguish between 'capacity' and 'utilisation'. Capacity, in the context of railways, refers to the maximum number of trains that a given section of railway infrastructure can handle over a certain period. Infrastructure utilisation, on the other hand, indicates how effectively this available capacity is being employed to deliver rail services. It encompasses not only the sheer volume of trains operating on the network, but also how well their schedules are adhered to, how evenly train services are distributed across the infrastructure, and how quickly and effectively disruptions can be managed.

3.2.2 Measures and Indicators

The ability to calculate capacity was deemed to be impossible by the International Union of Railways (UIC) in *UIC Code 406, 1st Edition* 2004 originally a capacity calculation method, citing "Capacity as such does not exist." but instead deeming that "Railway infrastructure capacity depends on the way it is utilised" (*UIC Code 406, 1st Edition* 2004). Although describing a method for its calculations it was stating that in its view a method to calculate absolute or true capacity was unlikely. Though the revised UIC 406 method (*UIC Code 406, 2nd Edition* 2013) no longer states this original view.

The method outlined in *UIC Code 406, 1st Edition* 2004, provides a systematic approach for assessing railway line capacity using the principle of timetable compression. Initially, a reference period is selected, typically spanning 24 hours, within which all scheduled trains are sequenced based on their designated paths. Subsequently, the timetable undergoes compression, positioning each train successively and eradicating any dormant intervals between them. Post-compression, the unoccupied time segments depict the railway line's residual capacity. The capacity utilisation, U, is then computed using the relation:

$$U(\%) = \left(1 - \frac{\text{Free time post-compression}}{\text{Total reference period}}\right) \times 100 \tag{1}$$

Interpreting this resultant percentage is crucial to discern the railway line's efficacy. As a guideline within the UIC 406 framework, a capacity utilisation exceeding 85% typically suggests that the railway is approaching its operational boundary, thereby possessing limited scope to reliably support additional traffic. Further investigations of this is studies using techniques similar to traffic flow and fundamental diagrams, such as in Corman and Henken 2022. Nonetheless, it is paramount to consider that tangible railway capacity can be swayed by myriad factors, inclusive of infrastructure quality, diversity of train services, and operational adaptability.

The UIC 406 method through timetable compression has been used as a standard practice for researchers in this field as a benchmark to refer to, extending on the original 4 parameters of "number of trains", "average speed", "stability" and "heterogeneity" for determining capacity (*UIC Code 406, 1st Edition* 2004) to include other features. For example those determined by Abril et al. 2008 in Figure 3.1.

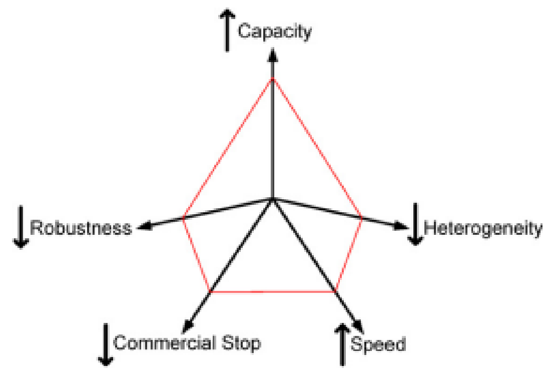


Figure 3.1: Conceptual view of the main parameters that affect capacity as determined by Abril et al. 2008 and how they vary from that in *UIC Code 406, 1st Edition* 2004 . (Abril et al. 2008)

3.2.3 Methodologies and Models

Research in the area typically highlight two main methods to calculate capacity, using analytical methods (such as timetable compression based on UIC406 (*UIC Code 406, 1st Edition* 2004; *UIC Code 406, 2nd Edition* 2013)) or employing simulations of static and stochastic models to determine the results. Murali et al. 2010; Pachl 2002; Sameni, M. Dingler, et al. 2011 used these groups with different research objectives. Some have suggested minor variations to these groupings to include optimisation (Abril et al. 2008; Mikulčić and Mlinarić 2021; Sameni and Moradi 2022). Though both Mikulčić and Mlinarić 2021; Sameni and Moradi 2022 added parametric and graphical methods to determine the capacity.

Determination of the capacity allows for the understanding of the utilisation, though the relationship between capacity and utilisation is debated.

3.2.4 Case Studies

A comprehensive review of case studies featured in existing literature concerning railway capacity calculation methodologies was performed. The aim was to gain insights into the geographical areas considered in these studies and the assessment methods utilised within them. In particular the global representation in railway capacity research, exploring the wide spectrum of locations from Europe and Asia to North America. Additionally, the distinctive network model used in the approach across these studies was evaluated, to better understand their application in different geographical contexts. A crucial part of the review involved analysing the extent of empirical data and real-world validation in these studies, recognising their importance in enhancing the reliability and practical relevance of the research findings. The aspiration was to gain a broad perspective on the state of railway capacity calculation methodologies in real systems and identify potential avenues for future research or methods.

The comprehensive review of the literature in Table 3.1 highlights an interesting mix of methodologies employed in the calculation of railway capacity. While some studies like Abril et al. 2008 and Azadi Moghaddam Arani, Jolai, and Nasiri 2019 have used empirical data and simulations for real-world validation, a substantial portion of the literature reviewed does not. Predominantly, the case studies cited in these research works span Europe, Asia, and North America with a prevalent use of modelling techniques, though with only one assessment of the UK railway (Sameni, Landex, and Preston 2011) from over a decade ago. Noticeably, there is a conspicuous absence of real-world validation across a significant portion of these studies, indicating a potential area for future research. The diversity of methods and geographical locations emphasised in these works underscores the complexities and nuances involved in railway capacity calculation, setting the stage for further in-depth exploration in this field. One area noted was the lack of use of geospatial data in the models themselves instead utilising a topographical model in almost all case studies.

3.2.5 Challenges and Solutions

The degree of infrastructure utilisation plays a pivotal role in service delivery, affecting the punctuality and reliability of train services. This, in turn, has a profound impact on customer satisfaction and operational costs. Moreover, effective utilisation can minimise the need for costly and disruptive infrastructure expansions by ensuring that the existing capacity is used to its full potential. Therefore, understanding and enhancing infrastructure utilisation is of utmost importance for the sustainability and resilience of railway operations.

The ideal method of determining capacity is concluded by Sameni and Moradi 2022 to be dependent on the available resources, constraints and aims/objectives of the desired investigation to draw on the strengths from the desired approach.

Its noted in reviews by Mikulčić and Mlinarić 2021; Sameni and Moradi 2022 that further research is needed into the analytical and simulation methods themselves, but with the intention of finding different ways to manage the railway in order to obtain the desired outcome from the results. A topic missing from these reviews but also in literature was a method to validate the outcomes or the capacity methods itself. Though some validation was provided, it was typically for homogeneous situations.

Table 3.1: A summary of literature examining various capacity methods and techniques assessing their use of empirical data in the study and if actual data was used in the validation of the methods. Additionally assessing the locations of case studies used in these methods and the type of network model used

Reference	Empirical Data	Practical Validation	Case Study	Continent	Model
Abril et al. 2008	Yes	Empirically	Spain	Europe	Topological Graph
Azadi Moghaddam Arani, Jolai, and Nasiri 2019	Yes	Simulation	Iran	Asia	Topological Graph
Boysen 2013	No	No	Sweden	Europe	Topological Graph
Burdett and Kozan 2006	No	Simulation	North America	North America	Topological Graph
Burdett and Kozan 2006	No	Simulation	Australia	Australasia	Topological Graph
Burdett 2015	No	No	North America	North America	Topological Graph
Burdett 2015	No	No	Australia	Australasia	Topological Graph
M. H. Dingler, Y.-C. Lai, and Barkan 2009	No	No	North America	North America	Topological Graph
Francesco, Gabriele, and Stefano 2016	No	No	Italy	Europe	No Model
Gavspark, Abramovic, and Hals 2015	No	No	Slovakia	Europe	Topological Graph
Goverde, Corman, and D'Ariano 2013	No	No	Netherlands	Europe	Topological Graph
Guo et al. 2016	No	No	China	Asia	Topological Graph
Harrod 2009	No	No	North America	North America	Topological Graph
Jamili 2018	No	No	Iran	Asia	Topological Graph
Jensen 2015	No	No	Denmark	Europe	Topological Graph
Jensen, Landex, et al. 2017	No	No	Denmark	Europe	Topological Graph
Jensen, Schmidt, and Nielsen 2020	No	No	Denmark	Europe	Topological Graph
Jia et al. 2021	No	No	China	Asia	Topological Graph
Y.-C. Lai, Liu, and T.-Y. Lin 2012	No	No	North America	North America	Topological Graph
Y.-C. Lai, Huang, and Chu 2014	No	No	North America	North America	Topological Graph
Y. C. Lai and Y. J. Lin 2016	No	No	Taiwan	Asia	Topological Graph
Landex 2009	No	No	Denmark	Europe	Topological Graph
Li et al. 2019	No	No	China	Asia	Topological Graph
Liao et al. 2021	No	No	China	Asia	Topological Graph
T.-Y. Lin, Y.-C. Lin, and Y.-C. ". Lai 2020	No	No	North America	North America	Topological Graph
Ljubaj and Mlinarić 2019	No	No	Croatia	Europe	Topological Graph
Mussone and Calvo 2013	No	No	Switzerland	Europe	Topological Graph
Pouryousef and Lautala 2015	No	No	North America	North America	Topological Graph
Rais et al. 2020	No	No	Malaysia	Asia	Topological Graph
Riejos et al. 2016	No	No	Spain	Europe	No Model
Sameni, Landex, and Preston 2011	No	No	England	Europe	Topological Graph
Sameni, Landex, and Preston 2011	No	No	Denmark	Europe	Topological Graph
Sameni, M. Dingler, et al. 2011	No	No	North America	North America	Topological Graph
Valentinovic and Sivilevicius 2014	No	No	Lithonia	Europe	Topological Graph
Wang, Nie, and Tan 2020	No	No	China	Asia	Topological Graph
Weik et al. 2020	No	No	Sweden	Europe	Topological Graph
Wu and Zhang 2019	No	No	China	Asia	Topological Graph
Zhong et al. 2019	No	No	China	Asia	Topological Graph

The literature review reveals that topological models, predominantly Node-Edge structures, form the backbone of railway capacity calculation methodologies as shown in Table 3.1. These models are typically either provided or self-generated, often relying on various linear referencing systems or commercial/proprietary format. However, this review has also highlighted a noticeable absence of geospatial technology application.

Capacity analysis generally involves the calculation of average speed, a key factor determined by the relation between distance and time. The accuracy of distance measurement, therefore, directly impacts the reliability of capacity calculations. It's interesting to note that while railways have their own distinct referencing systems for distance measurement, they could potentially benefit from the increased accuracy offered by geospatial data and globally comparable specific coordinate reference systems.

Implementing geospatial data, particularly when primary models are not available or when they are of subpar quality (as mentioned in the DARM (System Operator 2019)), could provide accurate distance measurements essential for reliable capacity calculations. Using open-source geospatial databases, like OpenStreetMap, could address this need and provide an additional benefit.

However, while these capacity methodologies provide a solid theoretical foundation, their effectiveness hinges on practical validation within real-world systems and the variations of a complex operational system. It's essential to examine these methodologies in context, assessing their applicability and accuracy within operational railway networks. This highlights an important research opportunity—exploring practical validation methods that can effectively test these methodologies, ensuring their relevance and reliability within real-world settings.

Thus, future research could benefit significantly from integrating geospatial technology into railway capacity calculations and focusing on practical validation to assess the real-world applicability of these methodologies. This would not only enhance the accuracy and reliability of these calculations but also foster global collaboration and knowledge exchange among railway infrastructure managers.

4 Methodology

A logical methodology, depicted in Figure 4.1, was derived based on the capability to construct the geospatial spatiotemporal graph of each service. Both planned and actual paths were considered, due to practical real world variations. This approach facilitates subsequent calculations related to the utilisation and capacity of the railway network. Various methodologies were investigated to ascertain the spatiotemporal graph with the utmost accuracy and precision, especially in light of data quality concerns.

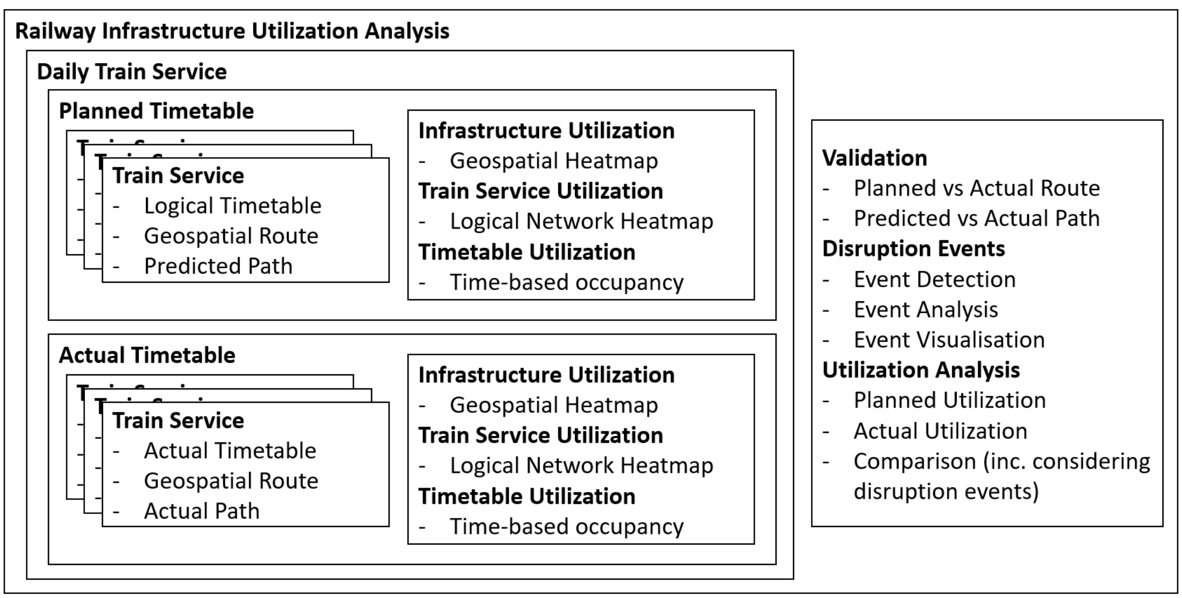


Figure 4.1: An overview of the defined methodology process designed to fulfil the aims and objectives of the investigation. This is achieved by calculating outputs for each train according to both planned and actual timetables, followed by an extended analysis across the entire network and train service.

The practical implementation of this methodology necessitates several components, each responsible for calculating the desired steps for individual trains and overall analysis. Additionally, the methodology accounts for the compiled datasets or models required for these calculations. This process is detailed in Figure 4.2. The key sections of these processes are elaborated upon in the subsequent subsections of this methodology.

4.1 Data Sources & Processing

The capacity and utilisation of the railway network depend directly on the quality and availability of pertinent data to determine these metrics. An absence of microscopic data, such as individual train positional information, makes it challenging to derive microscopic spatio-temporal details from available macroscopic sources. These sources are analysed based on the infrastructure and monitoring systems outlined in Section 2.3 and the example from Welwyn Viaduct mentioned in Section 2.4. The methodology should consider the breadth of the investigation and manage the required data's substantial significance and volume. A detailed summary of the datasets used can be found in Appendix Table A.1.

4.1.1 Timetable & Realtime Train Processing

A typical daily timetable for the UK Railway encompasses around 35,000 trains, peaking at a concurrency of 2,000 trains. There are approximately 1 million train movement events, as exemplified in Table 2.2. This timetable is issued daily around 01:45 and needs processing by 04:00, in time for the commencement of daily services.

Further complicating matters, there are 2.75 million TD (Section 2.3.4) step events daily spanning over 50,000 signals, presenting a considerable scalability challenge. An

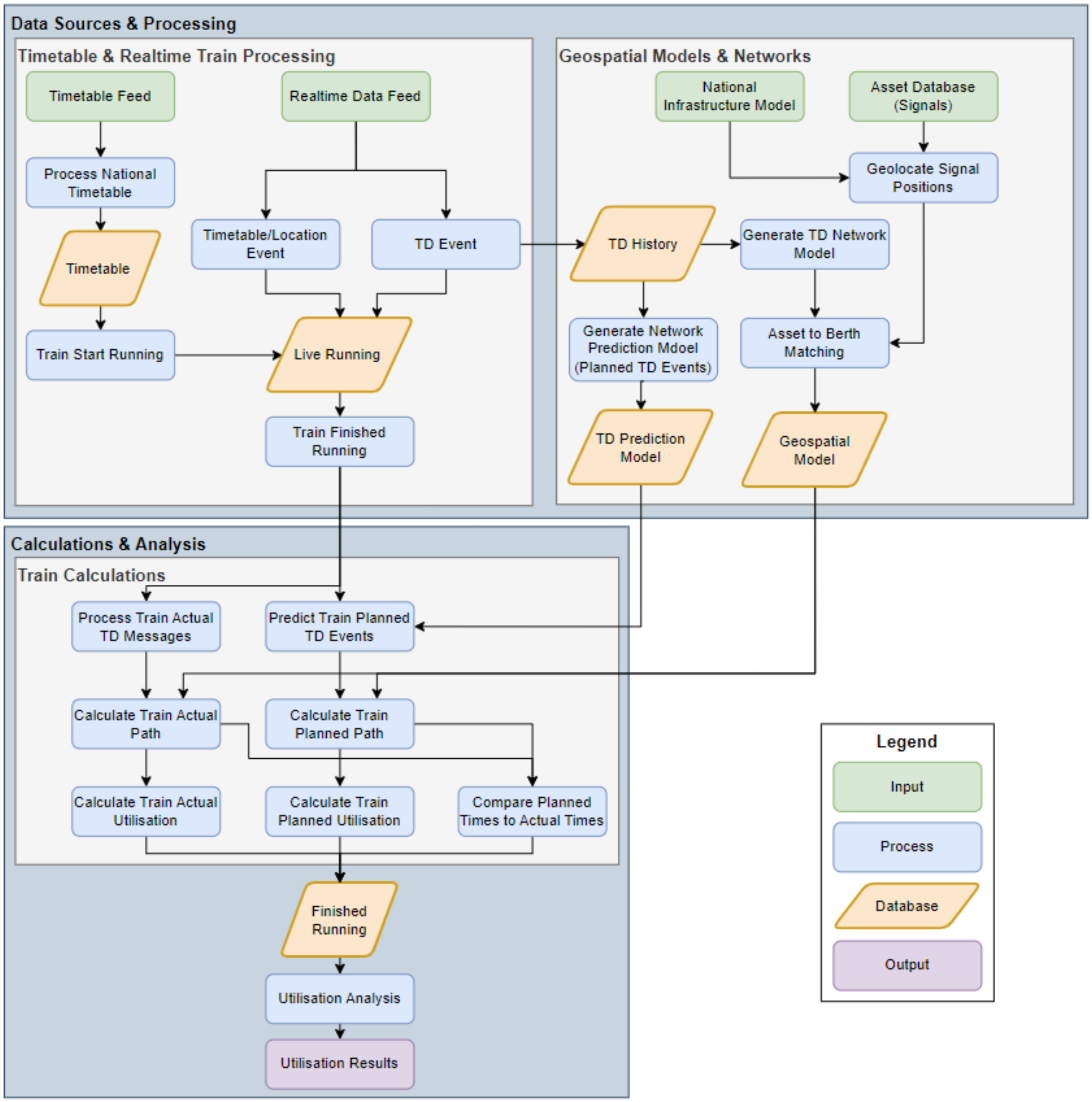


Figure 4.2: A detailed flowchart illustrating the process to calculate the methodology as presented in Figure 4.1. Data inputs are sourced and then processed. The left side of the flowchart showcases the processing of real-time train data and the processing for each train service upon completion. The upper right corner highlights the computations involved in the various models utilised in the calculations. It's noteworthy to mention the connection from live data (TD Events to TD History) which is used as historical training data in the prediction model.

illustration of this data is in Table 2.3. This intricacy is particularly noticeable when plotting paths across the vast railway network spanning more than 20,000 miles of track.

Due to industry legacy issues, the headcodes (four-character alpha-numerical Train IDs) aren't unique within the UK railway network. This results in complexities in deducing messages and assigning them to the appropriate services. For instance, "2J22", the most prevalent headcode, is scheduled for 14 distinct services every weekday as of May 2023. Occasionally, over five of these services can operate concurrently, demanding significant computational resources to determine the specific service to which an event pertains.

Owing to time limitations, not all system reliability events were integrated into the data processing. These events ensure that the timetable remains current with all modifications. Events excluded from the system are:

- Very Short Term Plans - about 900 services effected a day, 3.00% of daily services
- Train Reinstatement - about 60 services effected a day, 0.22% of daily services
- Change of Origin - about 215 services effected a day, 0.74% of daily services
- Change of Identity - about 40 services effected a day, 0.14% of daily services
- Change of Location - about 50 services effected a day, 0.17% of daily services

These counts are averaged over 219 days from November 2022 to May 2023, inclusive. In sum, these changes would impact 4.27% of all services, with an average of 29,879 services daily over this period.

4.1.2 Geospatial Model

The geospatial model's development synthesises various data sources identified while examining the physical infrastructure and the logical systems controlling the railway, as detailed in Section 2.3.

While the data from these sources is widely accepted (with some references to data quality and assurance in the DARM (System Operator 2019)), anomalies in data values are apparent but generally minimal. However, when consolidating these models, certain challenges were observed, necessitating corrective measures or elimination of the desired relational integrity between the sources. Two primary concerns arose:

- Bridging the gap between the linear and geospatial referencing systems.
- Matching a physical asset with its logical identity in the absence of a common reference system and the presence of non-unique elements.

The upcoming subsections will delve into these issues and potential solutions or workarounds. Discrepancies in source data within the train processing (Section 4.2) can influence the results, although such inconsistencies are addressed to the best extent possible in these procedures. For instance, discrepancies might manifest as implausible or invalid routes resulting from a signal associated with an incorrect line (e.g. the Slow line instead of the Fast line).

4.1.3 Geolocation of Signals

To geolocate signals, a process that interpolates signals and their linearly referenced positions into a geospatial referencing system is employed. The linear referencing position is specified in miles and yards from a specified datum. The NIM model, made publicly accessible by Network Rail through Freedom of Information requests, is presented in the OSGB36 coordinate reference system (EPSG:27700).

Signal assets come with identifiers for specific tracks, akin to the metadata found in the NIM model (refer to Figure 2.4 for an example). This facilitates deriving a geospatial position by interpolating the signal along the geospatial line string, proportionally based on the linear reference positional values. This procedure is depicted in Figure 4.3.

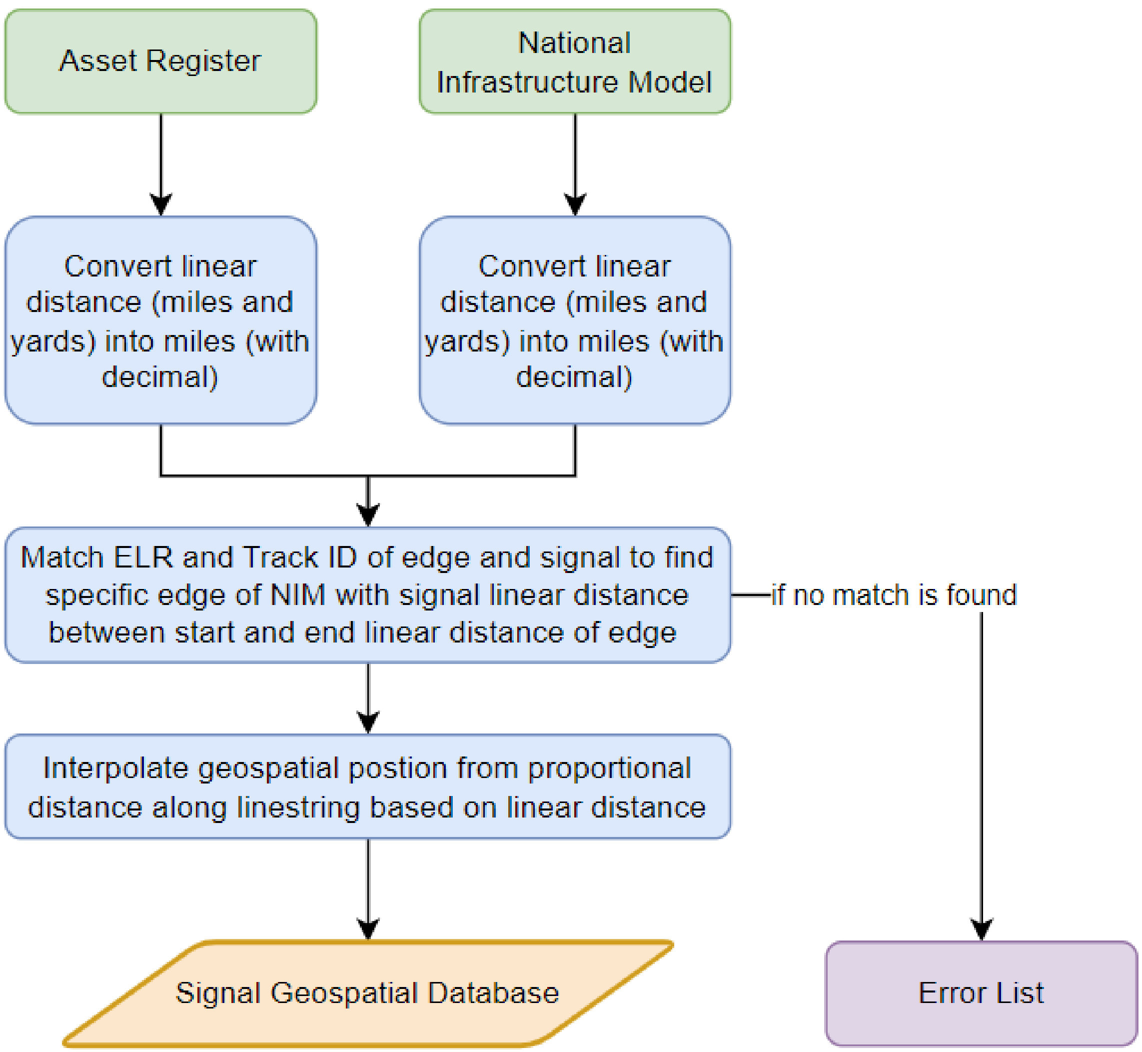


Figure 4.3: An overview of the process used to geolocate the signal position from a linear referencing system to a geospatial referencing system in the NIM.

However, this approach is vulnerable to data quality issues. For example, an incorrectly associated track ID in the source data could lead to an erroneous or absent reference, or referring to the wrong datum might result in positional inaccuracies.

4.1.4 Match Signals to Berths

To synchronise the layers between the logical monitoring systems (that provide train locations via Train Describers) and the physical assets (which have a geospatial location once geolocated), it's essential to match the signal asset ID to the berth ID.

While some similarities exist between the Signal ID and the Berth ID, as shown in Figure 2.3, this is not consistent throughout the UK network. Also, as evident in Figure 2.7, a single physical signal can correspond to multiple berths, either in cases of transitions between TDs or at major stations. The opposite is also true where a single berth can correspond to multiple signals.

A list of unique berths was generated by identifying distinctive berths in the TD historical events over a certain period. This yielded a compilation of potential berths to pair with signal assets.

The process of matching these two identities can be intricate. Figure 4.4 provides an overview of the process, aiming to establish a dependable method allowing a train's geospatial position to be identified when moving between signals (berth step). A straightforward approach was adopted to match these based on their numerical signal/berth numbers. If a unique match was elusive, further exploration using the SMART database and an additional location dataset (depicting geospatial positions of railway locations like stations) was done to create spatial polygons for each area ID (train describer). The nearest polygons were probed to increase the likelihood of finding a unique match.

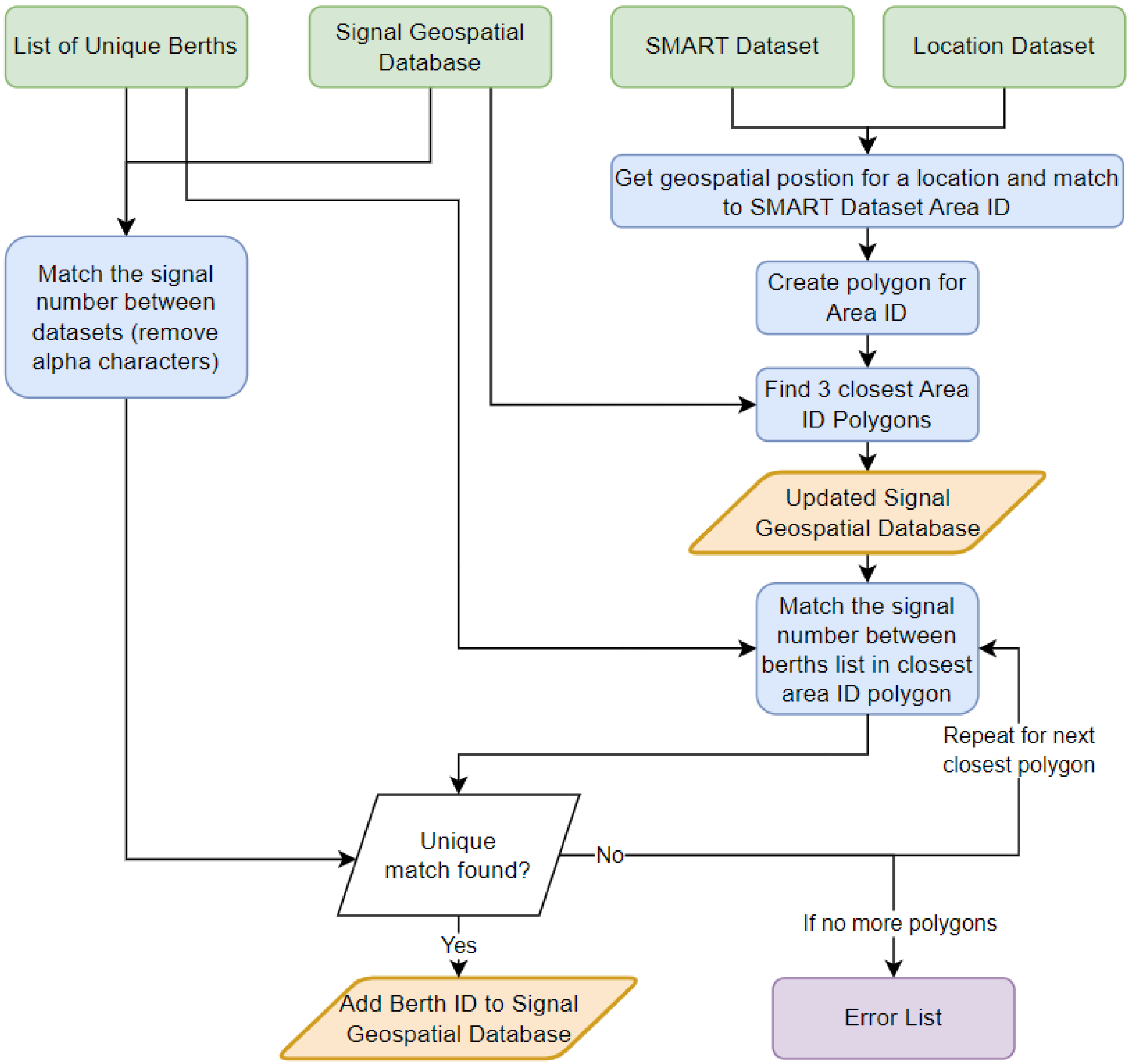


Figure 4.4: An overview of the process to align signal identities with berth identities, facilitating the assignment of a geospatial position from the signal's geospatial database to logical locations in the train describer monitoring systems.

While several methods were explored, this technique proved most effective, yielding the most matches. However, this method is not without its challenges, mainly due to the inconsistent nomenclatures between train describers. For instance, some signals (in the case of duplicates or multiple berths for a single signal) might receive additional alpha characters and use the last two digits of the signal id. This inconsistency means that local differences cannot always be taken into account.

4.1.5 Path Prediction Model

As outlined in Section 3.1, there are various methodologies, from Markov Chains to advanced neural networks, to determine a train's planned path through the TD networks. Graph Neural Networks methods however were not investigated as part of this project due to time constraints.

A robust historical dataset, which spans over 8 months and includes more than 300 million records of TD Events, serves as the foundation for developing an effective prediction model. Given that most train services run on an hourly frequency and are repeated weekly or even daily, this dataset is particularly repetitive and rich. For the prediction model, two labels were generated for each TD event: the subsequent step the train would take and the time interval between events. This labelling helps not only in predicting the path but also in integrating a temporal element to the graph, which can then be juxtaposed against the planned timetable.

Considering the base timetable's biannual changes, data from the last change (21st May 2023 to 23rd July 2023) was exclusively used, narrowing down the dataset to 112,673,689 events.

The initial task involved deciding the structural representation of the TD network, which could be either a flattened network or connected smaller networks. Figure 4.5 visualises these concepts, while Figure 4.6 demonstrates how a path might traverse both structures.

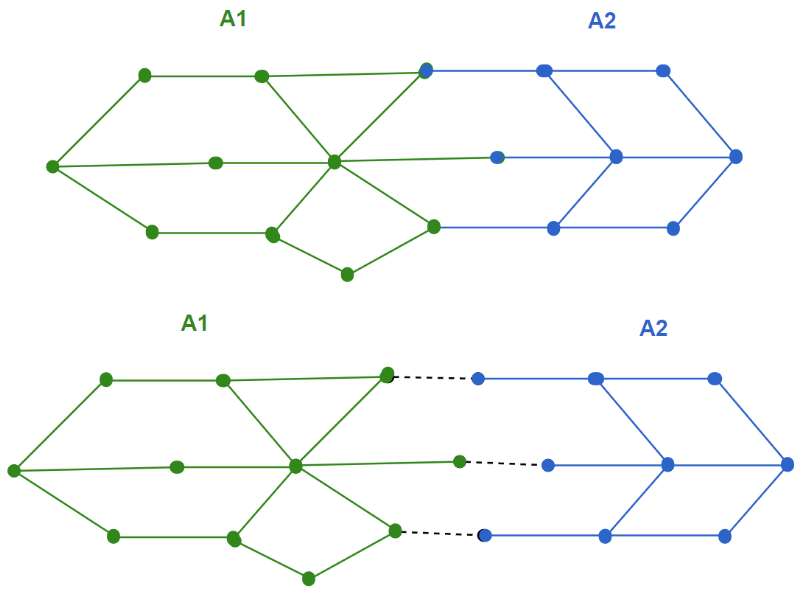


Figure 4.5: An example of a flattened TD network (top) and a connected TD network (bottom) where a single flattened network merges the TD networks (A1 and A2) together as a single network, top network, in contrast to the bottom network that connects separate TD networks (A1 and A2) together through inferred links (black dotted lines)

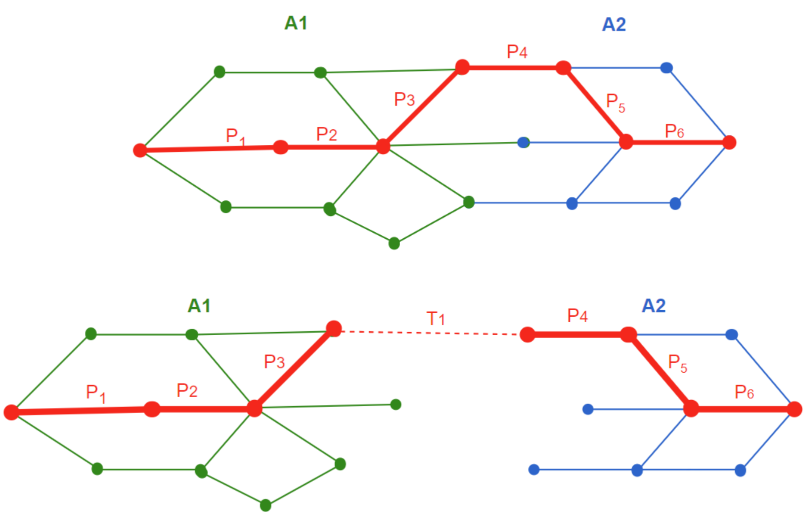


Figure 4.6: Example of the path (P_n) taken over both example network architectures described in Figure 4.5, a transition between the two networks in the connected network are shown by Transitions (T_1).

Both architectures come with their advantages and limitations. A flattened network, while simplifying connections between TDs, enlarges the network significantly. On the other hand, connected networks maintain network compactness but increase the overall number of separate networks to 181 (TDs). It was chosen to use the connected networks to prioritise the accuracy of the prediction by keeping the networks small and curate a list of transitions between the networks from the historical data.

Several techniques from 3.1 were explored. Traditional network path computations, such as Dijkstra's or BFS/DFS, could calculate paths between nodes but might not always reflect a train's intended path, especially given the implicit timetable concept highlighted in 2.3.1.

Neural network approaches were also explored, drawing parallels to next-token prediction in advanced language models. Despite efforts to adapt LSTM models for this task, the outcomes were subpar, often predicting impossible next steps. A shift towards training on a specific TD area network showed an improvement, but accuracy remained below 50% even within a small TD network.

Consequently, a simpler methodology was adopted. Given the recurring nature of timetables, a train is likely to adhere to its planned path. Leveraging historical data to identify the most common subsequent steps for each train emerged as an efficient predictive tool. This straightforward probabilistic model was scalable to vast datasets and can seamlessly integrated into a continual training pipeline. Its prediction accuracy exceeded 90% in initial testing when implemented in the connected network structure. The entire process is illustrated in Figure 4.7.

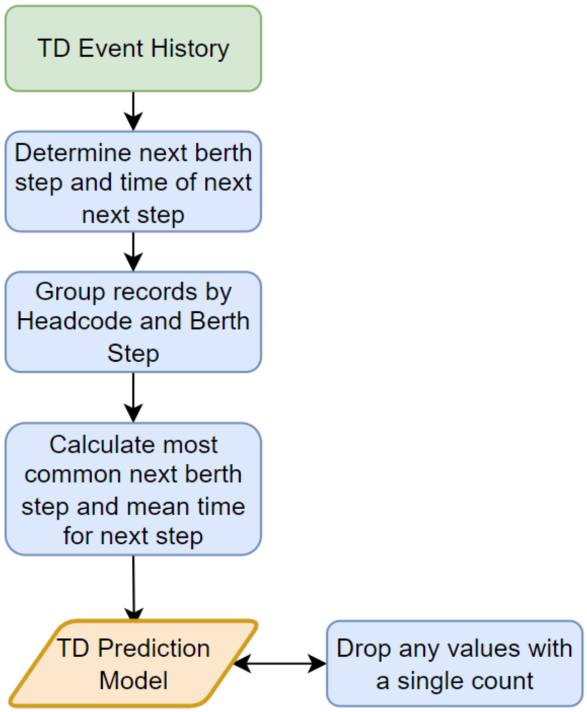


Figure 4.7: An overview of the process to create the path prediction model for predicting the path of TD steps through a network.

4.1.6 Path Prediction Inference

The prediction model will simply predict the next step and duration of that step when provided with an initial TD step. The initial step will be used as the starting location of a train is variable and once the train begins with the first step the remainder of the path can be calculated. The model will chain together these predictions to form a list of steps and duration's using the previous inferred step which will allow the model to truly predict the path of the train from start to finish only using the first actual path step as an input. This is visualised in Figure 4.8. The times will also be chained starting with the original departure timestamp from the trains timetable and then using the mean values from the prediction model.

Given the first actual step (A_1), the prediction model will chain together predic-

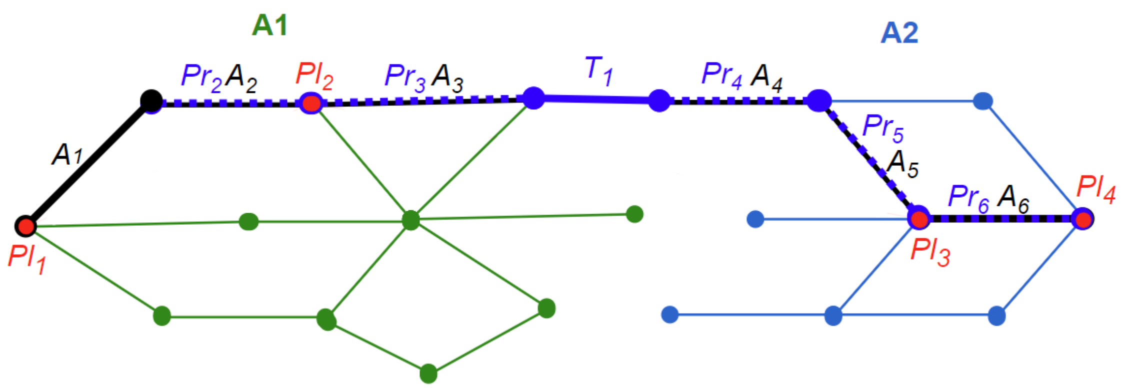


Figure 4.8: An overview of the prediction methodology to predict the path through the network. A1 and A2 are the connected networks by network transitions (e.g. T_n). The graph shows the actual path of the train (A_n) as bold black edges and planned timetable stops at specific points on the network (nodes Pl_n) as red, note that the planned timetable locations are at a lower resolution of positional accuracy with fewer nodes. The prediction is shown in blue ($Pr_{2...n}$). Where lines overlap dashed lines are used.

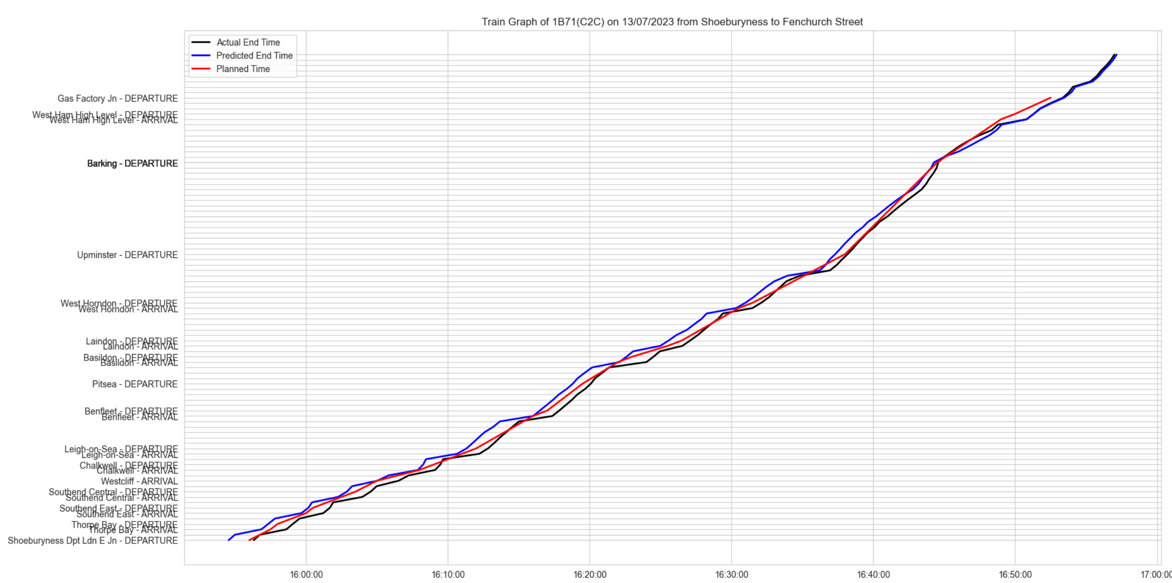


Figure 4.9: An example train graph were a real train (1B17 from Shoeburyness to Fenchurch Street (C2C Trains) on 17th July 2023). A train graph is a distance time graph with time against logical locations (Planned locations arrival/departure as per the timetable on the Y axis are named, however the unnamed ticks reflect steps in the network.

tions ($Pr_{2...n}$) using its previous step as the next predicted step. If the model were to reach an end point in one of the connected networks (e.g. A1), it would check to see if there was an available transition link (T_1) available to then use as the starting point of the next step and repeating the prediction element until the final point was reached or a point in the final area (e.g. A2) as this the train may arrive at a different platform than originally intended which is common practice.

4.1.7 Path Prediction Model Validation

Following on from the review of current literature the comparison of models to empirical data and validating the model against previous data can be advantageous to validating the performance of the model.

The results of the prediction model will be compared to and validated against two different data sets due to the different resolutions available and compounded effect of chains of predictions. This is seen in Figure 4.8 as the planned timetable (nodes Pl_n) and the actual path taken through the network (A_n) in the network.

The path prediction will be treated as a next token prediction with a binary classification where the next predicted step in the path is predicted correctly or not. Treating the prediction as this kind of problem would suggest the use of standard prediction metrics such as Accuracy, Precision, Recall and F1-score to be used as detailed below.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total predictions}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision will be most useful in this case of comparing to actual path data as the cost of a false positive on the prediction would be significant (predicting the wrong path). While recall and F1 may also be useful metrics detecting a true false prediction can be difficult when comparing to actual data (as the actual may not be a reflection of the plan) and as such will not be used in the results as it will be difficult to determine a False positive as the planned data at the step level does not exist.

The comparison to the planned location will still be relevant to assess the ability of the prediction to chain the correct sequence together and compare to the true planned path but at a lower resolution. Due to only being able to detect positive predictions and not actually predicting the planned data itself, only precision will be used when comparing the predictions to the planned locations as the number of correct predictions cannot actually be determined, rather only true positives can be detected where the planned location values align to the location nodes.

This is best visualised in Figure 4.9 which incorporates the temporal element of the graph. These are shown in accompanying colours to the example in Figure 4.8 (black for actual, blue for prediction and red for planned).

4.2 Calculations & Analysis

Assigning the correct and pertinent data to each train service is paramount, encompassing both the intended timetable and real-time performance metrics. The procedures needed to delineate both the envisioned and actual geospatial trajectories of train services are expounded in Figure 4.1, further detailed in the process chart of Figure 4.2.

4.2.1 Train Calculations

To discern a train's precise route and usage, the relevant event data for the specific service must be correctly attributed. The geospatial path is determined by locating the geospatial position of each TD step and subsequently determining the shortest Dijkstra path between signals (points) on the NIM model. Such computations yield the most detailed path of each train, in contrast to the inferred path via the timetable or origin to destination shortest-path calculations.

By harnessing the list of TD steps, the geospatial path across the NIM model's infrastructure network edges is computed, as exemplified in Figure 2.3. Subsequently, a consolidated geospatial path, inclusive of distance metrics utilising the projected coordinate reference system, is derived. Data resilience mechanisms are integrated; for instance, if a signal's geolocation is ambiguous or it can't be mapped accurately, a fallback mechanism involving the shortest path to the next relevant signal is employed. Supplementary logic ensures both the path's accuracy and the flagging of potential data inconsistencies.

Data sources encompass the real-time feed, inclusive of both the envisaged and actual timetables, as well as real-time TD events. Forecasted TD events are formulated based on the TD prediction models, enabling a harmonised path calculation mechanism for both sets of data, ensuring consistency and optimising processing efficiencies.

4.2.2 Utilisation Analysis

Given the intricacies of implementing certain capacity computations referenced in Section 3.2 within this project's duration, a more rudimentary approach was adopted. Infrastructure utilisation within the NIM model was gauged by a simple frequency count of each component's usage. For instance, if a train's deduced path overlapped with an NIM model component, its usage count was incremented. Over a given duration, network utilisation could be assessed by tallying the frequency of infrastructure component traversal in both the planned and actual timetables.

While simplified this method would allow for a relevant relational metric to be inferred, though implementing other capacity calculation methods would be a key part of future work(Section 8.

4.2.3 Calculations & Analysis Validation

As there is no recent existing work or figures available to validate the utilisation nor geospatial path of train services it will be difficult to validate the results of the analysis to existing empirical data, particularly on the scale and broad depth of geographical and infrastructure variations. Further work will be needed to validate the results over visual inspection of a sample of results which was used in this analysis.

4.3 System Implementation

To methodically capture and analyse this data, the methodology was implemented within a realtime system to constantly process the incoming data. Owing to the extensive scale and scalability requirements of the computations, the system was designed with cloud-native concepts in mind, oriented towards facilitating further investigations. The system design was implemented within the Amazon Web Services (AWS) cloud environment, and a depiction of the System Architecture can be seen in Figure 4.10.

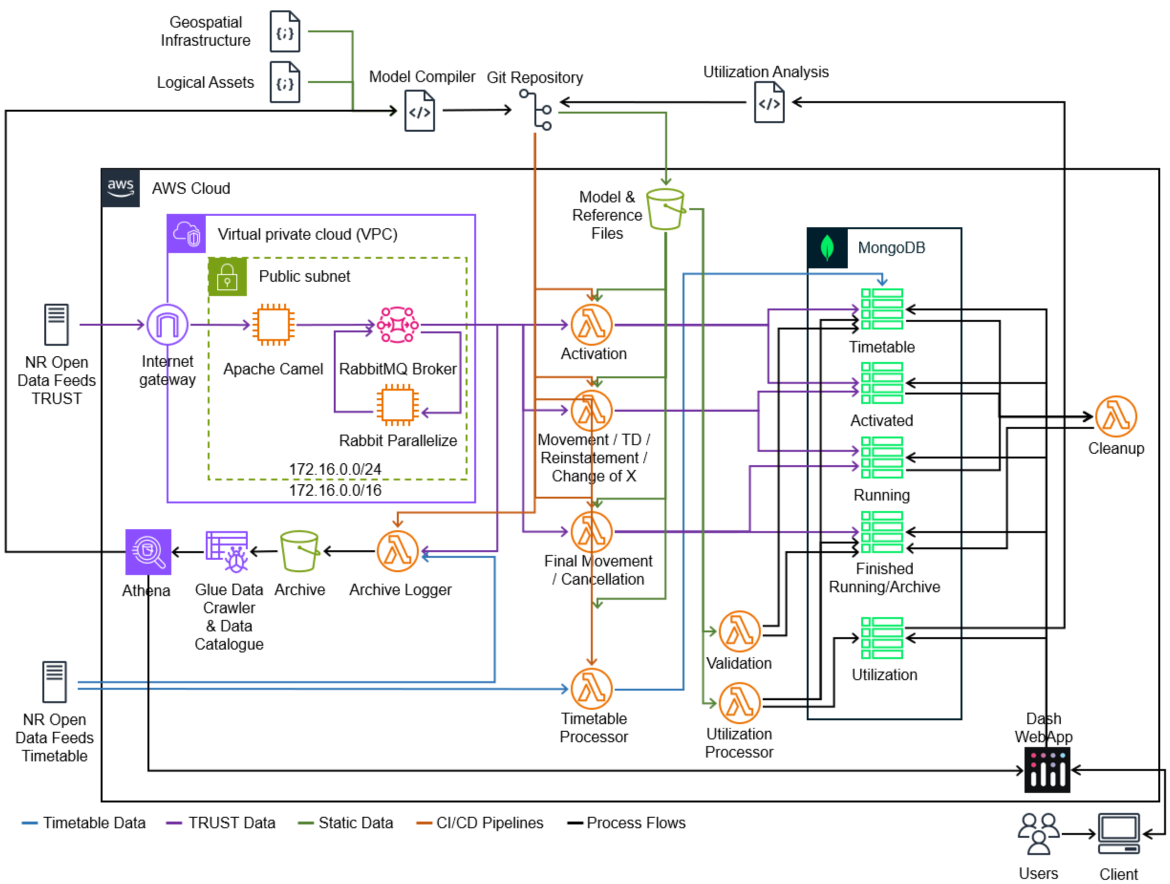


Figure 4.10: An overview of the intended realtime system architecture to implement the desired processes in Figure 4.2. The system has linkages to allow for reproducible analysis with updated data and ability to self generate and maintain data to improve system reliability and continual improvement

4.3.1 System Architecture

The proposed architecture (as seen in Figure 4.10) employs a NoSQL JSON document-based database, where each train is represented as a document. An Online Transaction Processing (OLTP) concept is used to ascertain the current state of the train service, this is the implementation of the Timetable, Live Running and Finished Running datasets where processes transmit a transaction (train service) between various states keeping the size of the database low but performing large complex calculations between states. This choice of OLTP over a conventional relational database is intended to enhance scalability and optimise data access patterns for processes that are computationally intensive. However, to characterise the derived layered model, as illustrated by the Geospatial Model or the TD History, Online Analytical Processing (OLAP) would be necessary to conduct analytical analysis and historical training/analysis of the prediction model.

The system has some ability to self-maintain and characterise the models over time, as recorded messages are stored and processed as part of analytical queries to update values or models, in addition to manual CRUD (Create, Read, Update, Delete) adjustments.

This leads to a mixed architecture that leverages the advantages of both OLTP and OLAP concepts, resulting in a self-maintaining data architecture, critical to help address the data quality concerns.

5 Results

The results section provides a comprehensive overview of the data collected and analysed during the course of this study. Rooted in rigorous methods and computational techniques, the findings shed light on both the predicted and actual utilisation patterns across the UK railway networks.

5.1 Data Acquisition and Processing Period Overview

The methodology outlined in Figure 4.1 and implementation of this in Figure 4.10 was run for a period from 11th July 2023 to 23rd July 2023. A summary of the quantities of data processed is shown in Table 5.1.

Due to a processing error, the total counts of timetabled trains for 14th, 15th, 17th, and 19th July 2023 were unable to be recorded. These are indicated on the table as N/A values. Additionally, on 20th and 22nd July 2023, there was industrial strike action which significantly affected the number of trains running on the network as detected by the system. This resulted in 8 weekdays, one Saturday, two Sundays, and two strike days of service. During this 13-day period, all trains on the UK rail network were processed and monitored. In total, 26,505,742 TD events and 7,795,620 train movement messages were processed from the realtime data feeds to enrich the timetable data.

During this period, an average of 32,546 trains per day were timetabled (this includes cancelled trains). On a normal weekday, a mean of 16,706 trains completed their journeys; of this, an average of 4.91% of trains failed to complete their journey once starting. Despite this, 82.55% of train services that completed their journey were able to have their predicted path calculated fully, with an average of 13,797 normal weekday service trains able to fully process all steps of the geospatial path and utilisation process for both planned and actual paths including utilisation calculation.

Table 5.1: A summary table of the Data Acquisition and Data Processing results for the collection period between 11th July 2023 and 23rd July 2023. Days where industrial strike action effected the train service are shown in **bold**.

Day	Day of Week	Timetabled	Completed	Incomplete	Predicted	Predict Incomplete
11/07/2023	Tuesday	33,145	17,677	854	14,588	3,943
12/07/2023	Wednesday	33,865	17,914	893	14,854	3,953
13/07/2023	Thursday	33,905	17,854	910	14,724	4,040
14/07/2023	Friday	N/A	17,816	912	14,703	4,025
15/07/2023	Saturday	N/A	14,303	788	11,726	3,365
16/07/2023	Sunday	20,677	9,459	386	7,644	2,201
17/07/2023	Monday	N/A	15,844	784	13,078	3,550
18/07/2023	Tuesday	37,040	15,939	777	13,102	3,614
19/07/2023	Wednesday	N/A	15,954	831	13,249	3,536
20/07/2023	**Thursday**	**39,243**	**7,634**	**406**	**6,590**	**1,450**
21/07/2023	Friday	39,176	14,654	776	12,079	3,351
22/07/2023	**Saturday**	**34,854**	**6,923**	**313**	**5,781**	**1,455**
23/07/2023	Sunday	21,011	9,210	368	7,355	2,223

5.2 Prediction Model

The prediction model was a probabilistic model of 2,934,452 rows with at least 2 samples for each headcodes/next step combination, which could lead to the creation of a planned path for a train through the railway network.

5.2.1 Prediction Model Results

The prediction model executed a total of 7,216,663 inferences for 190,179 trains. These were matched to 4,326,137 actual step events, resulting in a 62.94% prediction match rate to the actual data.

Table 5.2 provides an overview of the results both in the comparisons between actual step data (A_n) and planned timetable (Pl_n) data with the predicted values (P_n) inline with Sections 4.1.6 and 4.1.7. The table has 3 distinct sections: first, the total actual, predictions, and true positive prediction counts; second, the average values for all predictions; and finally, the average values considered against each train. The notation used is inline with Figure 4.8.

The prediction output results summarised for each train can be found referenced in the Appendix to the project data repository (B)

Table 5.2: A table the prediction model metrics for predicted trains and the prediction results. The table includes 3 main sections; initially total counts of actual values, predictions and true positive predictions, average total prediction accuracy and precision values and average counts, accuracy and precision for train averages. As it wasn't possible to calculate the predicted planned data the count and following accuracy values were unable to calculate, these are shown as N/A

Prediction Model Metric	Actual Step Data (A_n)	Actual Planned Data (Pl_n)
Total Count of Actual Values	8,208,714	3,796,307
Total Count of Predicted Values	7,216,663	N/A
Total Count of True Positive Predictions	4,326,137	1,744,179
Average Total Prediction Accuracy	0.60	N/A
Average Total Prediction Precision	0.53	0.46
Average Train Count of Actual Values	65.86	30.46
Average Train Count of Predicted Values	57.90	13.99
Average Train Prediction Accuracy	0.61	N/A
Average Train Prediction Precision	0.54	0.44

The model predicted train paths that had the highest number of steps – more than 700 steps. Out of these, 5 of the 6 services with over 700 steps were the Aberdeen to Penzance CrossCountry service (1V60), which operated on 11th, 12th, 13th, 14th, and 18th of July 2023 during the data collection period. Although the number of predicted steps for each service hovered just below 200 (with counts of 199, 196, 198, 199, 196 respectively), these predictions had an average accuracy of 0.18. The precision was even lower, at 0.05.

5.2.2 Precision Details

A total of 2,584 trains achieved an actual data precision value of 1 (perfect train path prediction). The Merseyrail Liverpool Central to Ormskirk service emerged as the most common, with 659 out of its 716 services obtaining this high precision value. With the exception of service 1B20 (South Western) on 16th July 2023 traveling from ELGH to WATRLMN, all services belonged to the same connected network. Notably, 1B20 traversed four distinct connected networks: BE, EH, SU, and WI. Table 5.3 shows the number of services with a perfect train path prediction of at least 10 occurrences.

5.2.3 Prediction Algorithm Errors

Unfortunately, 40,706 trains could not complete their predicted path from start to finish. This shortcoming can be traced back to several factors, primarily classifiable under three major types. These three types are based on the prediction model calculation exit error codes. These errors affected 27.74% of the train predictions. A list of these issues is showcased in Table 5.4. The causative factors for the unsuccessful train predictions include:

- Model Error - Situations where the model failed to reach the final node (or network). This could be due to a missing transition between networks or encountering an unconnected node within the prediction model networks.

Table 5.3: A table of the most common services by operating company with a count greater than 10 and an Actual Prediction Precision of 1. Locations are shown as TIPLOCs (7 letter code for a location) for From-To.

Opereating Company	Service (From-To)	Count
Transport for Wales	PENARTH-BARGOED	109
	RADYR-CORYTON	59
	PENARTH-RHYMNEY	54
	YSTRADM-PENARTH	44
	CORYTON-RADYR	23
	CRPHLY-CRDFCEN	19
	RHYMNEY-PENARTH	16
	BARRYIS-ABDARE	12
C2C	SHBRYNS-FENCHRS	54
	GRAYS-FENCHRS	26
East Midlands	CLTHRPS-BRTHMBR	44
Great Western	DIDCOTP-OXFD	67
	OXFD-DIDCOTP	33
	BEDYN-RDNGSTN	13
Greater Anglia	CLCHSTR-WONNAZE	95
	THPLESK-WONNAZE	46
	WONNAZE-THPLESK	39
	HERTFDE-BROXBRN	15
	LOWSTFT-NRCH	12
Merseyrail	LVRPLCH-ORMSKRK	659
Northern	BRADFS-SKPT	244
	SKPT-BRADFS	25
	COLNE-PRST	24
	GTSHDMC-NWCSTLE	12
Southeastern	STNGBRN-SHRNSOS	16
ScotRail	NWTL-GLGC	135
	ABRDEEN-INVURIE	98
	EKILBRD-GLGC	71
	MONTRSE-INVURIE	48
	GLGC-NEILSTN	26
	GLGC-EKILBRD	18
South Western	WOKING-WATRLMN	11
Nexus	AIRP-SNDRLND	12
Crossrail	CHDWHTT-ILFEMUD	18

- Prediction Error - Cases where the model could access the final node (or network) but an invalid piece of data triggered an erroneous prediction. An evident symptom of this was when a train exhibited a precision greater than 1.

- Infinite Loop - Instances where the prediction algorithm gets ensnared in a circular loop across the connected networks or within a network, rendering it unable to locate the final node (or at least final network).

Table 5.4: A table of error types and counts of the trains that encountered the error.

Error Type	Count of Error	% of Total Predicted Trains
Model Error	34,489	18.14%
Prediction Error	12,053	6.34%
Infinite Loop	6,217	3.27%

5.3 Utilisation Calculations

Building upon the discussions in Section 4.2.2, the utilisation metric was computed as a straightforward count of each edge's usage in the NIM model, derived from both the planned and actual spatiotemporal graphs of each train.

Computing the temporal graph necessitated extensive processing to assemble the requisite geospatial model, which in turn enabled path calculation from the planned and actual steps.

5.3.1 Utilisation Calculations - Geospatial Model

Achieving a precise alignment between physical signal sets and their spatial coordinates, and subsequently pairing them with the logical berth identity, was pivotal in formulating the geospatial model described in Section 4.1. Of the 40,651 signal assets, 2,312 (amounting to 5.69%) could not be accurately geolocated based on the linear-referenced positions relative to the NIM. Out of the successfully geolocated 38,249 signals, 17,434 were paired accurately with an appropriate berth identity. More than 15,000 of the remaining 20,815 signals managed at least a partial match but failed to secure a unique pairing. Visual representations of these outcomes can be viewed in Figure 5.1. Figures 5.2 and 5.3 offer snapshots of successfully matched bert identities, with varying degrees of accuracy. Adjacent signals, barring those at boundaries, should ideally share the same Area ID (initial 2 characters) to signify the TD network. This is evident in Figure 5.2, which displays 4 signals from Figure 2.7. However, this consistency, in line with Tobler's First Law of Geography, is conspicuously absent in Figure 5.3, where multiple Area IDs coexist within a compact geographical zone.

This discrepancy is further illustrated in Figure 2.4. Comparing the signals on this geospatial display with those in Figure 2.3 reveals the absence of signals like Y8_0628 and Y8_0629 from the geospatial perspective due to unsuccessful matches.

The elements of the compiled geospatial model can be found referenced in the Appendix to the project data repository (B)

Such misalignments impair the path calculation process. Instead of a seamless, linear trajectory akin to a train's movement, the path might evolve into a convoluted trail with improbable routes, like reversing over a set of points devoid of a logical or tangible path. This distortion in the path calculation skews the utilisation count, erroneously inflated counts for infrastructure segments untouched by any train. Although measures were integrated to compensate for missing signals, the repercussions of these discrepancies are evident in the utilisation results. Examples of faulty routing can be

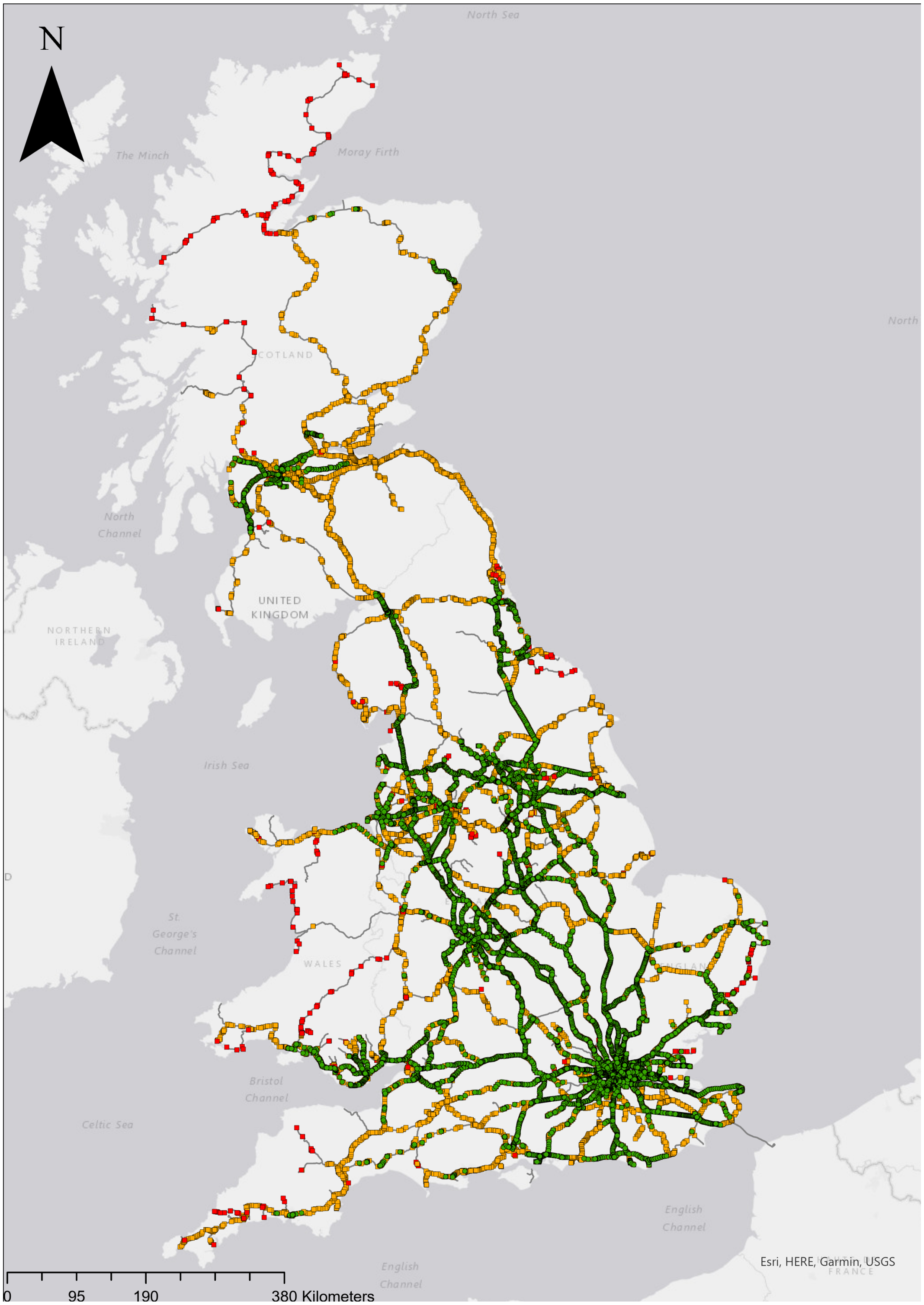


Figure 5.1: Overview of the Matching Results in the UK with Geolocated signals to their Berth.

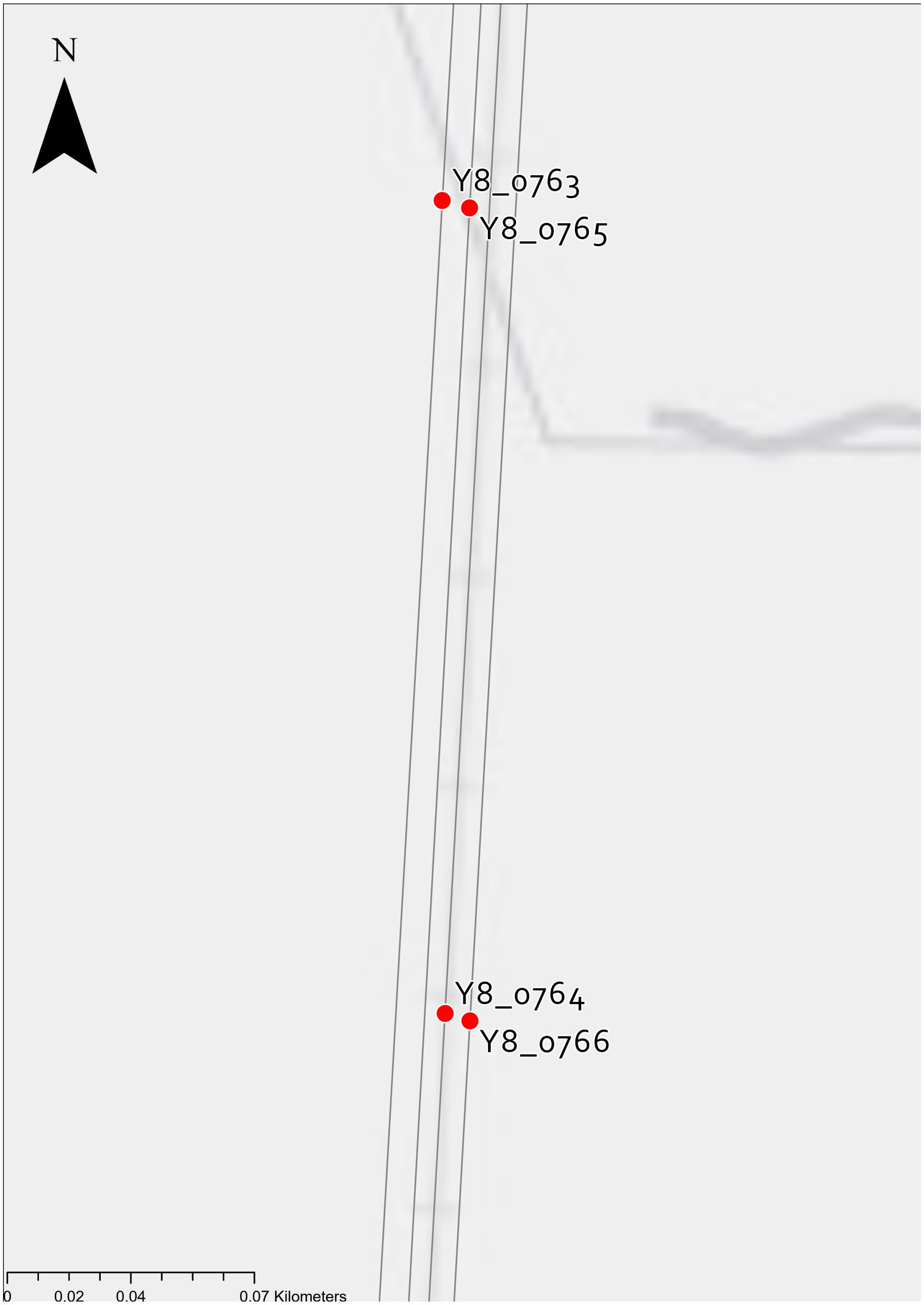


Figure 5.2: Example of Berth Allocation to signals around the area of Sandy in Befordshire, this a geospatial representation of some signals seen in Figure 2.7.

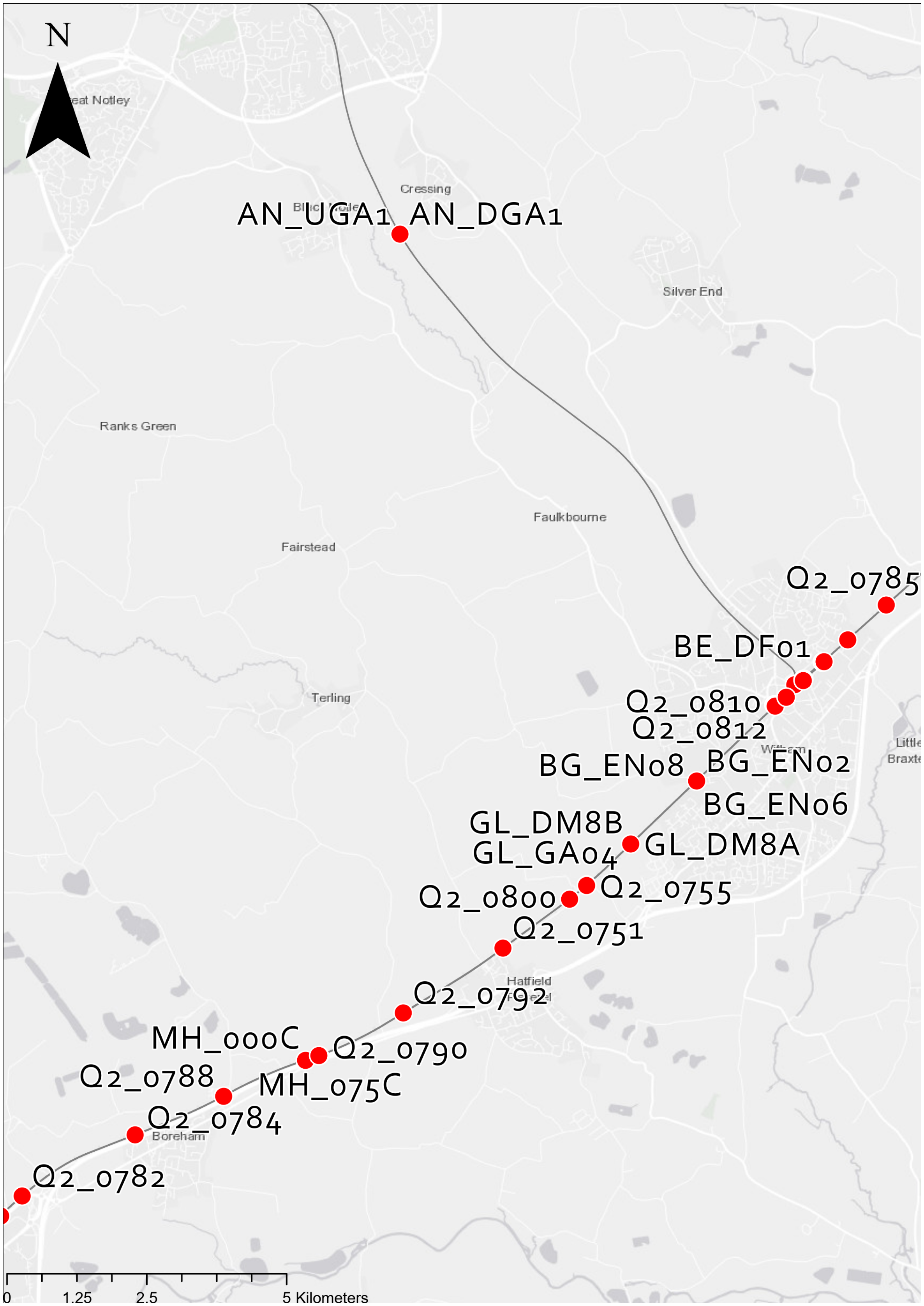


Figure 5.3: Example of Berth Allocation to signals around the area of Witham in Essex.

seen in Figure 5.5, especially when juxtaposed against Figure 5.4. In the case of Figure 5.5, the erroneous route to Basingstoke is a consequence of a berth match gone awry. Conversely, the route to Ipswich stems from a flaw in the source data's linear reference position, leading to a signal's inaccurate geolocation. This is merely a representative example of a successful prediction and route plagued with subpar results.

Quantifying the number of trains affected by these anomalies proved infeasible within the project's time frame.

5.3.2 Utilisation Calculations - Analysis

During the data collection period, both actual and planned utilisation was determined for processed trains, resulting in 149,173 trains being examined over a span of 13 days, as detailed in 5.2.1.

The computation required robust infrastructure, utilising the ARM-based 'cg6.16xlarge' virtual machine on AWS. This machine, with 64 vCPUs and 128 GB RAM, ran at 99.98% for over 7 days, reflecting the computational demands of the analysis for the 13 days of data collection.

Utilisation maps, computed on a daily basis, exhibited the differences between planned and actual paths. Comprehensive maps for the entire period were also produced, as depicted in Figures 5.7, 5.6 and a difference between these two total counts in Figure 5.8.

Each days normalised planned, actual and difference between the utilisation calculation can be found in the Appendix (Section C). Normalised Planned Utilisation Maps for the data collection period from 11th July 2023 to 23rd July 2023 are shown

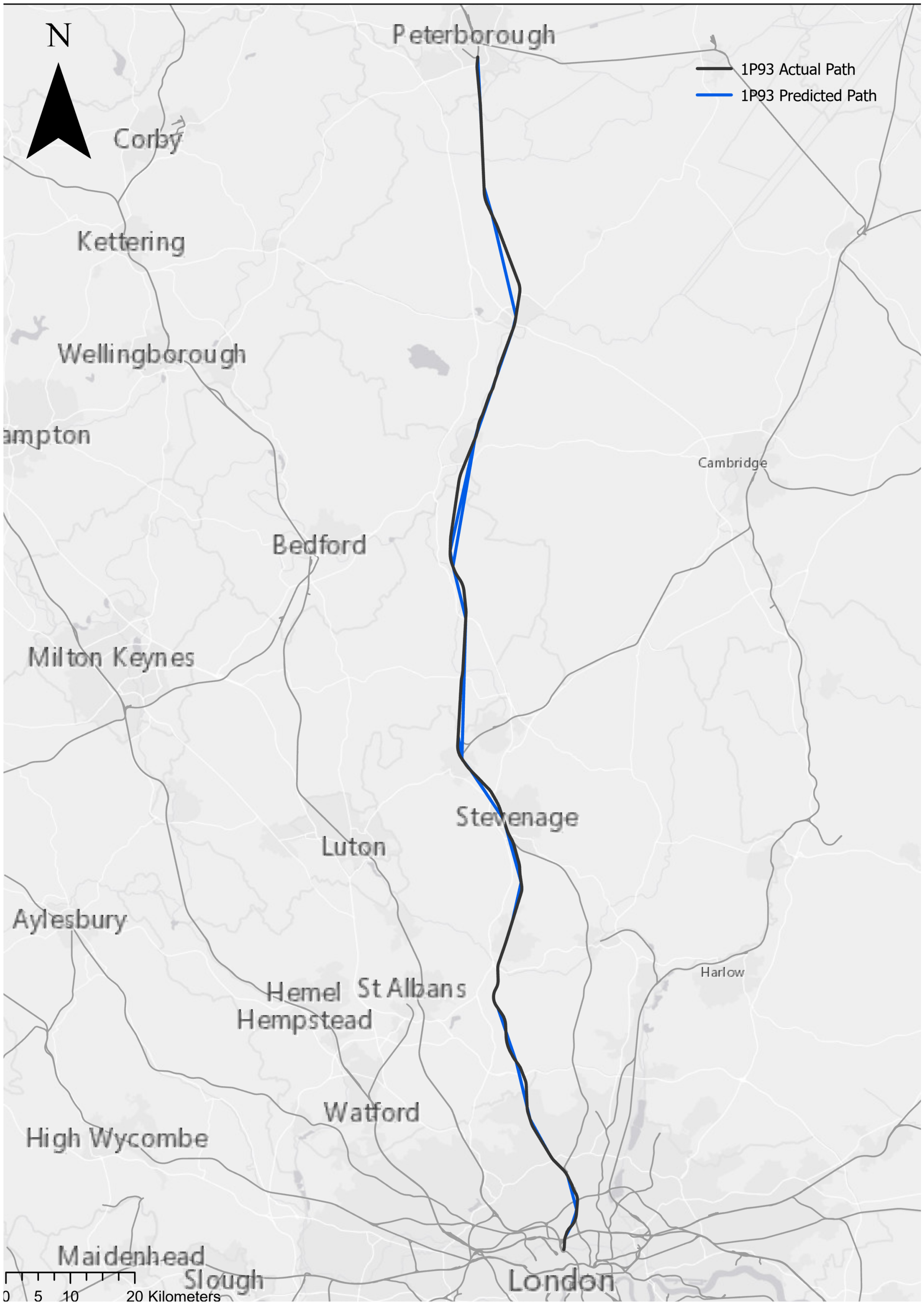


Figure 5.4: Example of good path calculations as seen for 1P93 (Peterborough to Kings Cross - X02209 on 21/07/2023), the same example used in Figure 2.4. While there are some divergences from the intended line there are some errors in the path calculation for the predicted path in particular.

Figure 5.5: Example of bad path calculations for 1Y08 (Sunderland to Kings Cross service - G23693 on 18/07/2023) both the actual and predicted path have significant incorrect routes with diversions to Ipswich and Leeds for the actual path, as well as diversions to Basingstoke and Woking for the predicted & actual path. Routing such as this cause incorrect utilisation values for the infrastructure off the original path.

in Figures C.1 to C.13 respectively. Normalised Actual Utilisation Maps for the data collection period from 11th July 2023 to 23rd July 2023 are shown in Figures C.14 to C.26 respectively. The difference between the Planned and Actual Utilisation Maps for the data collection period from 11th July 2023 to 23rd July 2023 are shown in Figures C.27 to C.39 respectively. The utilisation output values can be found referenced in the Appendix to the project data repository (B)

Inspection of these figures reveals regions with zero planned or actual utilisation, consistent with zones that lacked geolocated signal matches shown in 5.1. Major routes, especially the East, West, and Midland Mainlines, generally displayed heightened actual counts. Conversely, minor lines showcased elevated planned utilisation figures.

Histograms, as shown in Figure 5.9, suggest a potential log-normal distribution for utilisation counts. To verify this, a Shapiro-Wilk test was implemented on a sample of 5,000, chosen from the 15,416 utilised infrastructure elements. The outcome, presented in Table 5.5, hints that the log utilisation data might not follow a log-normal distribution and is supported by the plots seen in the Q-Q plots of Figure 5.10.

Table 5.5: A table of results from the Shapiro-Wilk test for normality of the log distributions of utilisation counts

Test	Variable	W-statistic	P-value
Shapiro-Wilk	Log Planned Utilisation	0.9463	2.60e-39
Shapiro-Wilk	Log Actual Utilisation	0.9544	6.35e-37

5.3.3 Utilisation Calculations - Strike Days

Though unintentional at the time, the inclusion of two strike days in the data collection period have rendered some interesting results. The industrial strike action

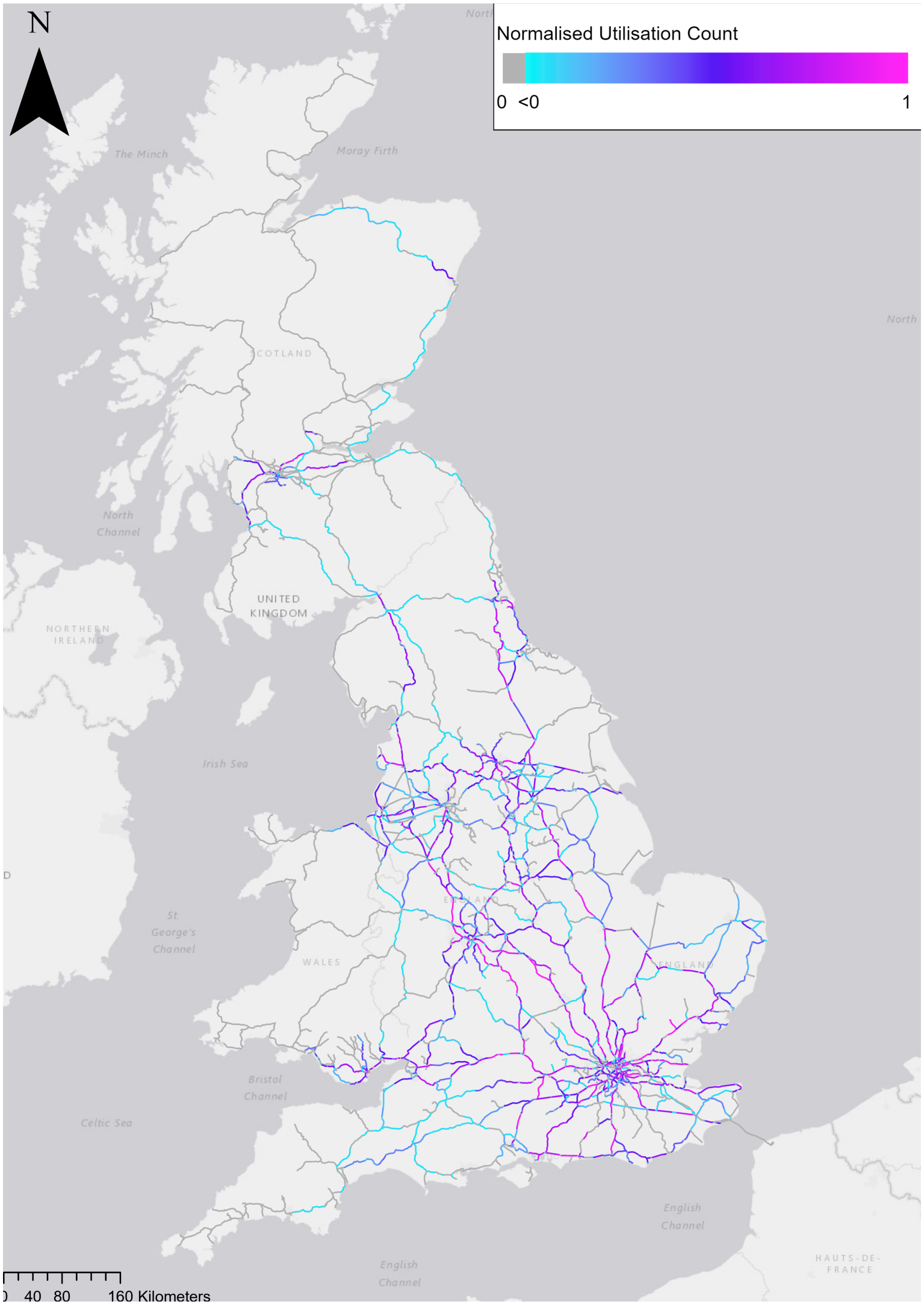


Figure 5.6: Map of the normalised utilisation from the actual train paths over the data collection period. A colour scale has been added to represent the utilisation from Cyan to Purple (low utilisation to high utilisation). Sections of the network with no utilisation have been greyed out.

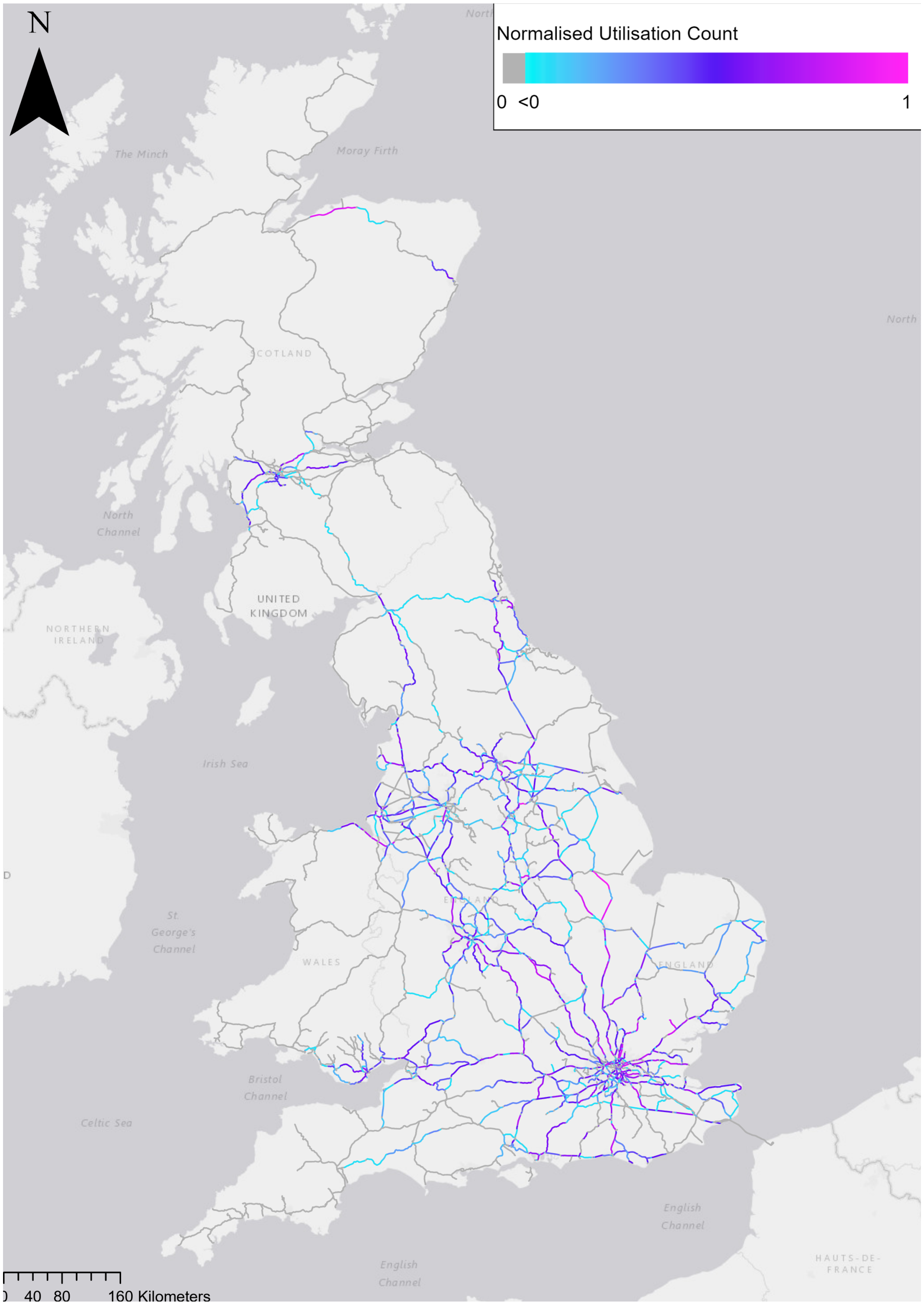


Figure 5.7: Map of the normalised utilisation from the planned train paths from the predicted path over the data collection period. A colour scale has been added to represent the utilisation from Cyan to Purple (low utilisation to high utilisation). Sections of the network with no utilisation have been greyed out.

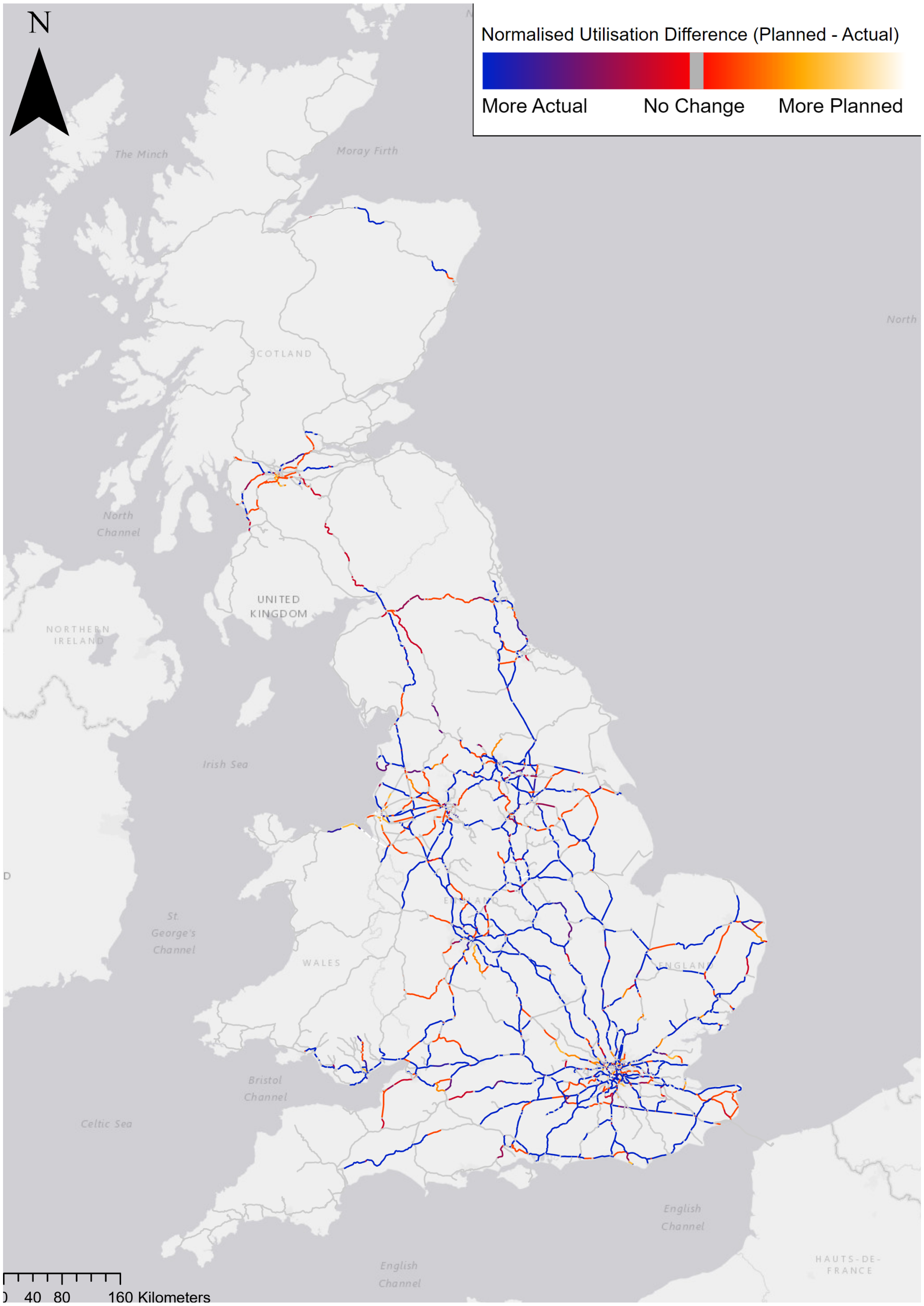


Figure 5.8: Map of the difference in total utilisation between the planned and actual train paths shown in (Figure 5.7 and Figure 5.6). A normalised colour scale has been added to represent the utilisation from Blue to Red to White (Increased Actual Utilisation to No change to Increase in Predicted Utilisation). Sections of the network with no difference have been greyed out.

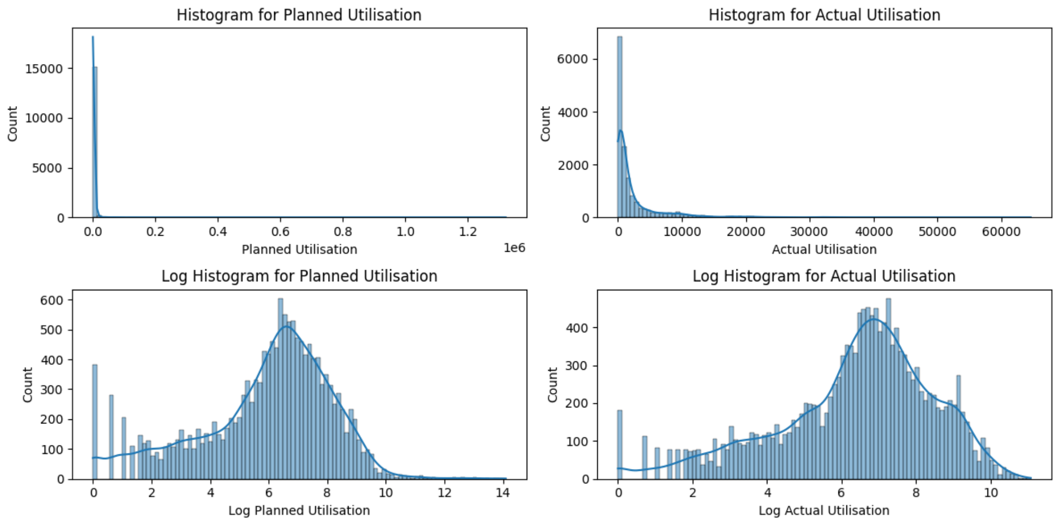


Figure 5.9: 4 graphs presenting the histograms of the utilisation values for the planned utilisation (left side) and actual utilisation (right side). The top graphs show the histograms for the counts of use of the elements of the infrastructure. The bottom graphs show the histograms of the log counts. Elements of the infrastructure with no utilisation as seen in grey in Figures 5.6 and 5.7 have been ommited from these.

by the RMT union involved 20 operating companies having their services significantly impacted. Comparing the relative difference between the number of operating company train services between the 19th and 20th July 2023 show the significantly effected services rise sharply to over 40% relative difference for those effected companies with Northern most effected at nearly 94.00%.

This can be further seen in the reduction in services run on the 20th and 22nd July 2023 in 5.1 with 7,634 and 6,923 services completed respectively. Focusing on the utilisation's on 20th July, as a greater difference in planned services due to a full week-day service, particularly the Merseyrail operating company as it operates a typically high density metro style service in the local areas of Liverpool with the regional areas supported by Transpennie Express, Northern, Transport for Wales and Avanti West Coast services.

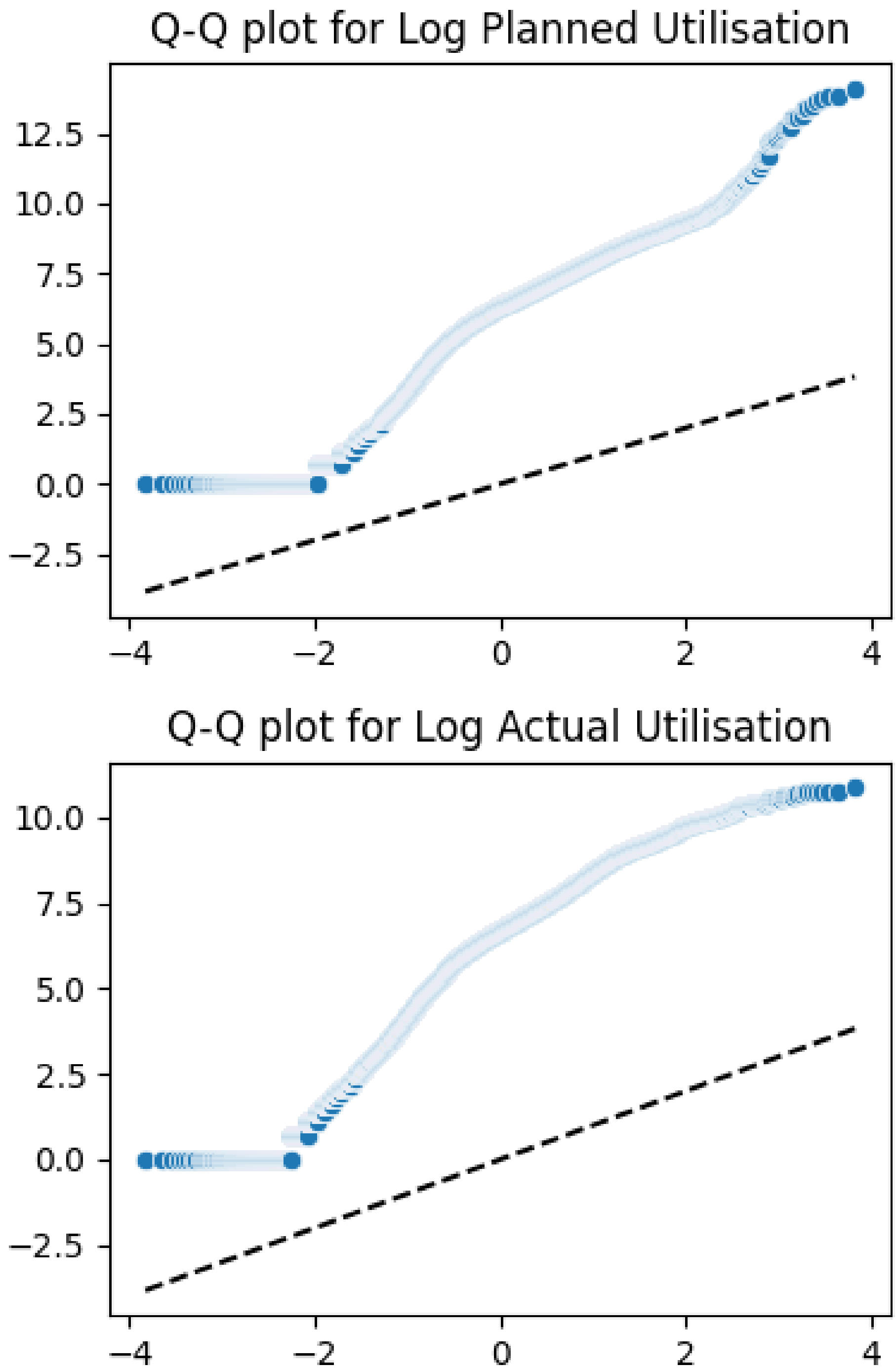


Figure 5.10: A Q-Q plot of the actual and planned log utilisation values which supports the results of the Shapiro-Wilk tests in Table 5.5 as the plots diverge from the diagonal line that would be expected of a log normal distribution

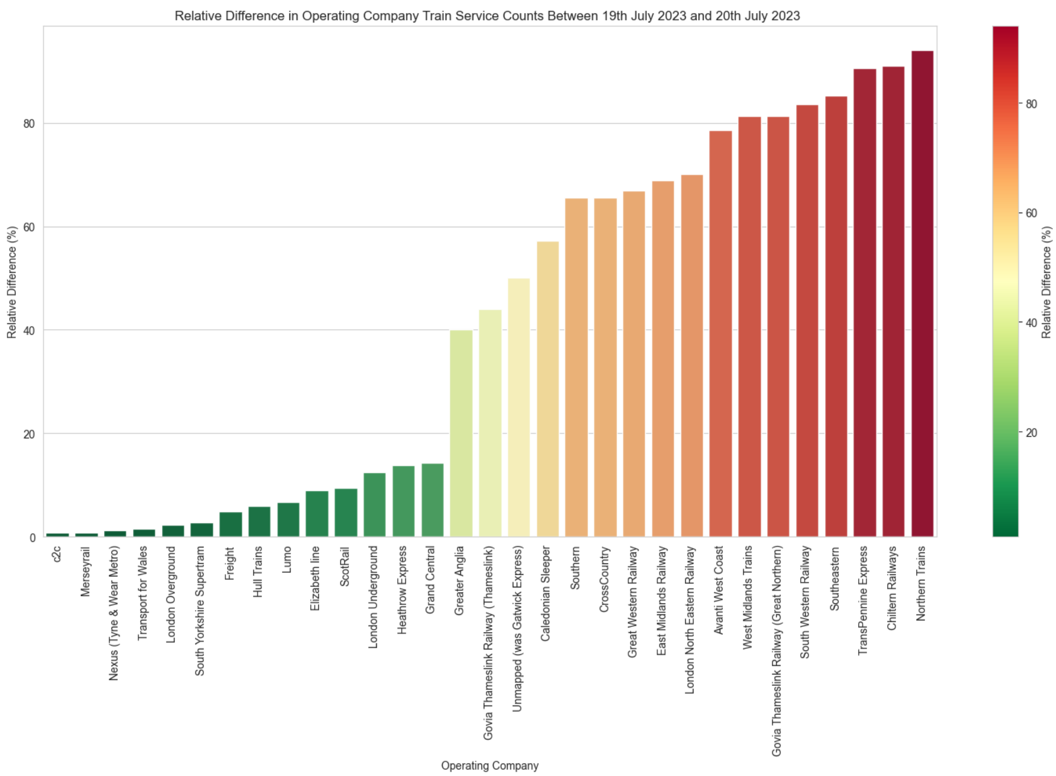


Figure 5.11: Relative differences in counts for various Operating Companies between 19th July 2023 and 20th July 2023. Most companies exhibit only minor day-to-day variances, such as c2c with a difference of only 0.82%. However, some companies have substantial changes over the one-day period. Notably, Northern Trains experienced a significant 94.00% reduction, followed closely by Chiltern Railways and TransPennine Express with drops of approximately 91%.

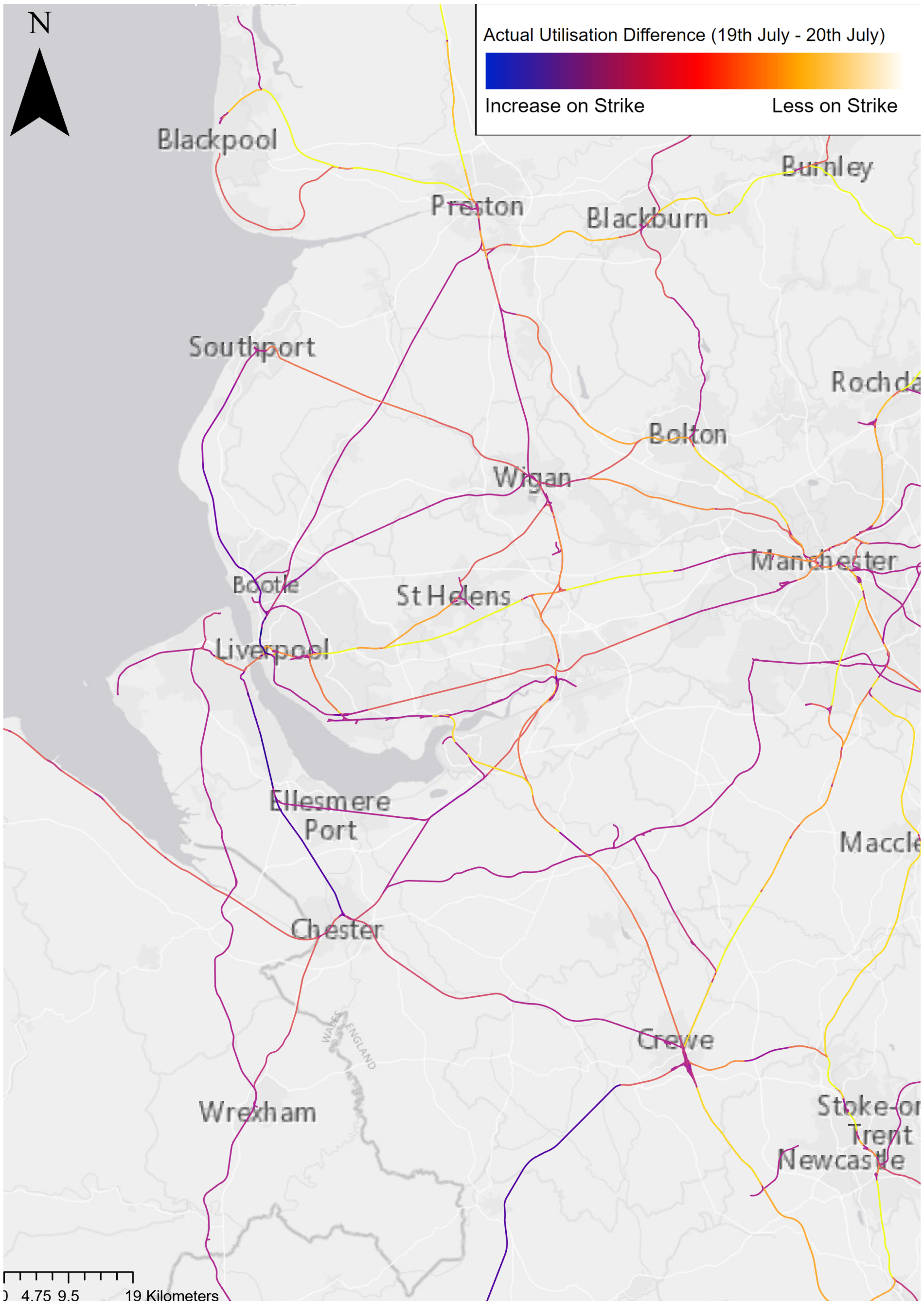


Figure 5.12: A map of the actual utilisation difference in North West England between 19th and 20th July 2023. The lighter the colour of the network indicates less services utilised the infrastructure on 20th July 2023 when the strike occurred as opposed to the 19th July 2023 which was a normal weekday service. Notable areas where less utilisation was seen is around Blackpool, Blackburn, Burnley and South of Manchester. These areas are typically served by Northern trains and Avanti West Coast. Darker colours can be seen around Liverpool and Chester/North Wales where Merseyrail and Transport for Wales operate.

Visual inspection of Figure 5.12 show a clear difference in the utilisation values represented by the lighter colours to show less services used those areas of the infrastructure. It supports the validation of the methodology where as seen in Figure 5.11 Merseyrail and Transport for Wales services running at a similar service correspond to darker areas in Figure 5.12. Where areas operated by Northern and Avanti West Coast around Blackpool, Blackburn and south of Manchester typically show a lighter areas indicating less use which aligns with the proportion of operating company train services run.

6 Discussion

The railway network's dynamic nature demands intricate analytical tools and methodologies for efficient and optimal planning. As the industry continues to evolve, so too must the strategies used to assess and predict utilisation, ensuring the most effective allocation of resources and understanding of systemic behaviours. This discussion delves into the various facets of our analytical approach, critically examining the intricacies of data quality, validation, anomalies and system implementation. Each segment sheds light on the successes, challenges, and potential avenues for further research and optimisation in our quest to create a comprehensive model for railway infrastructure utilisation and planning.

6.1 Prediction Model

The probabilistic model utilised for prediction, albeit simple, proved to be accurate. It showed a pronounced number of true positives, surpassing those observed in initial investigations of other models, including neural networks and transformers. Initial phases of the methodology development of these other methods yielded subpar results, indicating a need for further feature engineering within the geospatial model. The success of these adjustments would largely depend on the model's existing features and the quality of other data sources. As delineated in the methodology Section 4, and observed in the results, especially the utilisation values in Figures 5.7 and 5.6, the quality is currently lacking. Inaccuracies, such as those evident in Figure 5.5, are notably detrimental to this investigation.

Having access to a comprehensive historical dataset, like the one containing historical TD events, greatly influences the prediction model's performance. While the full dataset spanning 8 months (around 300 million events) wasn't employed, the data since the last major timetable change was utilised (comprising 112 million events). Investigating the impact of this choice on the probabilistic model's performance would be enlightening, possibly enhancing model quality as observed in Table 5.4. One limitation of the probabilistic model is its inability to handle new or unprecedented train services that aren't in its historical data.

The chosen connected network architecture for the prediction network, as illustrated in Figure 4.5, seems appropriate. Yet, as highlighted in Section 5.2.2, services that achieved perfect predictions were restricted to a singular connected network. The transition list, designed to bridge these networks, introduced certain complications, including infinite loops. These were mainly attributed to the extensive range of network boundary transitions. Lacking a predetermined transition list, a heuristic method was employed to establish a consistent list for effective data processing. Notable errors arose when long services could only generate predictions for limited sections, a case in point being the CrossCountry services between Aberdeen and Penzance (1V60), as discussed in 5.2.1.

A merit of the probabilistic prediction model lies in its potential for consistent behaviour. Contrary to a fuzzy logic based model (such as neural networks), iterative enhancements targeting aspects like dataset size or model architecture could offer a structured approach to find systematic improvements.

6.1.1 Validation

Validating the prediction model is intricate, particularly when juxtaposing planned versus actual data. While the prediction yields a planned outcome in the resolution of the actual data, validation can only be bench marked against the planned timetable data, which possesses a coarser macroscopic resolution.

Discrepancies between the planned and actual data, such as rerouted trains, are perceived as prediction errors. Yet, juxtaposed against the planned data, these might appear accurate but at a diminished resolution. Tracking predictions against planned timetables poses challenges since datasets don't always precisely pinpoint train locations. Confidence in accurate predictions arises from trains operating exclusively within one TD area, like the Merseyrail services between Liverpool Central and Ormskirk. However, this dimension was overlooked in the prediction results' computations, suggesting future analyses should incorporate this metric to further investigate and improve multi network performance. This introduces an additional complexity related to the transition between connected networks, impacting data validation as depicted in Figure 2.7 and Table 2.4. While these transitions might compromise validation scores, they simplify the transition process by offering a singular connection link and reflect actual and correct behaviour of the service as opposed to the recorded event data.

6.1.2 Temporal Component of the Prediction

Throughout the development of the probabilistic prediction model to determine the modal next step led to the integration of the step's mean duration into the model.

This facilitated the generation of spatial-temporal paths during the model inference, exemplified in Figure 4.9. This temporal feature, paired with the next-step prediction, simplified the task of comparing predictions against planned timetables, which inherently are spatial-temporal graphs (albeit of the timetable and planning topology rather than the geospatial infrastructure network). Although it wasn't the central aim, the temporal dimension of the prediction opened avenues for intriguing explorations and potential future work.

6.2 Routing Calculations

The geospatial path through the network was calculated for both the actual performance and planned timetable (using the predicted values). Both calculations were performed using the same functions, ensuring repeatability and comparability. Echoing the sentiments in Section 6.1, the probabilistic nature of the prediction model means that the routing calculation exhibits predictable behaviour. This allows for continuous method improvement by comparing outcomes from varying changes. Although this repeatability is beneficial, the diversity of situations, geographical complexities, and data quality introduces a myriad of scenarios. Thus, any methodological modifications might yield wide-ranging performance impacts.

The path calculation process attempted to offset the challenges posed by poor-quality data through input validation and iterative optimisations, aiming to exclude outliers but was not as successful as intended resulting in a number of invalid calculations. Additionally, the process had to account for missing data, and a balance had to be struck to achieve effective results.

Employing shortest path calculations over the network model, specifically between signals, proved beneficial. This approach negates the limitations of the implicit timetable methodology. As shown in Figure 5.4, this method yields accurate routing in areas with complex infrastructure. However, when juxtaposed with traditional routing methods used in traditional routing problems (eg SatNav Routing) like Djkijstas, DFS, or BFS (refer to 3.1.1) these may be more resilient to the anomalies seen, unless a specific waypoint is inaccurately placed as seen in this project.

Undoubtedly, the quality and availability of data play pivotal roles, not only in determining routing capabilities but also in influencing the most suitable method. Further experimentation is required to evaluate the benefits and ill effects of each method, and combining both might enhance calculation performance.

6.2.1 Routing Validation

Large-scale validation of the routing remains elusive; current validation relies on visual inspection of a sample of calculated paths. Given that routing is not tracked currently in the railway network, using microscopic data sources like GPS isn't viable due to inadequate data accuracy, especially in areas with multiple train lines. The desired GPS accuracy required is estimated to be less than 30cm to correctly determine the explicit track path. Though the use of future digital signalling technology in the European Train Control System (ETCS) would make this possible, but roll out of this technology is currently limited in the UK and access to the required microscopic or macroscopic data is not currently accessible.

6.3 Utilisation Analysis

The utilisation results described in 5.3 offer several points for further exploration. A pivotal factor influencing the results is the presence (or lack) of a precise and comprehensive geospatial model. This is evident in the issues observed in Figure 5.5 and in the results of the planned (Figure 5.7) and actual utilisation (Figure 5.6).

The employed metric to simply count how often an infrastructure element is traversed during train path calculations is very simple and directly subject to data quality issues. Notably, this metric is absent from the literature discussed in 3.2. This choice was made since the data required for more intricate metrics, such as the standard UIC 406 ((*UIC Code 406, 1st Edition* 2004; *UIC Code 406, 2nd Edition* 2013)) method, weren't readily available. While the geospatial approach can inherently compute some of these required features, like section length, obtaining a comprehensive dataset was challenging within the project's time frame and scope.

The supposed log-normal distribution of both actual and planned utilisation, as outlined in 5.3.2, where Figure 5.9 hints at such a relationship, is intriguing. However, the Shapiro-Wilk test (Table 5.5) and Q-Q plots (Figure 5.10) dismiss the idea of the data following a normal distribution. This outcome aligns with expectations, as no such relationship appears to exists, in line with conceptual expectations.

6.3.1 Data Quality

Using normalised values in both planned (Figure 5.7) and actual utilisation (Figure 5.6) allows meaningful comparisons between the two. However, given that only

15,416 out of 49,252 infrastructure elements from the NIM model are used by the prediction algorithm, room for improvement is evident, and not every train service is fully considered or calculated correctly. Areas shaded grey in Figures 5.7 and 5.6 correlate with predominant green signal matches in Figure 5.1. Drawing from the example in Section 5.3.1, it's evident that incomplete and imprecise signal-to-berth matching impacts the accuracy of utilisation calculations, though the impact and sensitivity of this is not known.

6.3.2 Validation

The project has yet to thoroughly validate the utilisation calculations, given the quality of the output results. A significant question remains, regardless of current data quality issues: How do we validate these outcomes? As this metric is uncharted and any data for its computation is currently absent, alternative validation methods are needed. Monitoring selected locations using lineside cameras might be a potential solution for smaller regions. Alternatively, a programmatic comparison against lower-resolution data, like the TD network, might offer a method of broad-scale and ongoing systematic validation. As seen in recent literature (Section3.2.4) the validation of capacity or utilisation findings is challenging and is rarely compared to empirical data or fully validated to real application.

6.3.3 Strike Days

While challenges persist in the utilisation outputs seen in Figures 5.7 and 5.6, insights from Section 5.3.3 hint at a potentially valid methodology. The observed differences in Figure 5.11 between train services on 19th and 20th July 2023 (strike day) in North West England suggest the methodology's correctness. This merits further investigation, as it suggests the methodology's potential. The unintended inclusion of the strike day in data collection, though unexpected yielded valuable insights, confirming the methodology's promise.

6.4 System Implementation

The system architecture, designed to execute the desired methodology on a vast geographical scale and national train service, proved effective. Utilising NoSQL databases like MongoDB and cloud-based processing capabilities, such as AWS Lambda, scaled to meet national network demands. Given the promising insights from the strike day analysis in Section 5.3.3, improving data quality and optimisation could pave the way for a feasible realtime system implementation.

Significant processing overhead remains, offering opportunities for added time sensitive computational steps without affecting the system's performance or stability. The data amassed to 111GB, roughly equating to 373,000 trains with an average document size of 597.77kB.

Potential improvements include decoupling prediction and routing calculations. Implementing a message broker or queuing system could yield asynchronous, parallel processing capabilities, facilitating more scalable computation. This would transition from a single, resource-heavy machine for Extract Transform Load (ETL) processing (5.3.2) to multiple smaller, cost-effective machines capable of real-time processing allowing for a more effective horizontal scaling capability.

However successful, the development faced hurdles, particularly with geospatial processing packages for python like GDAL and Fiona. Transitioning to serverless processing models may enhance maintainability and autonomous system administration. Yet, current performance and cost challenges persist. A shift to a precompiled routing model may address this, but the implications, especially concerning data quality, need careful examination.

7 Conclusion

The exploration of railway infrastructure utilisation has illuminated its potential and challenges. Our discussion has elucidated the complexities of routing calculations, stressing the imperative for consistent and comparable methodologies. The juxtaposition of actual and planned paths, buttressed by geospatial methodologies, signifies the strides technology has taken and its potential in refining railway systems. While our methodology provides key insights, its repeatability and comparability are paramount for systematic improvement.

The application of straightforward methodologies, such as the probabilistic prediction model, is feasible due to the railway system's intrinsic behaviours. Anticipating future enhancements with integrated complex concepts might yield superior results, contingent on the quality and availability of data.

Data quality emerged as both a foundation and an impediment. The divergence between planned and actual utilisation, amplified by data quality challenges, accentuates the essence of comprehensive data coverage. Creating an accurate geospatial multilayered model by achieving a holistic signal-to-berth match and encompassing all train services is crucial for improved precision and accuracy.

Regarding system implementation, the adoption of contemporary technologies like NoSQL databases and serverless computational offerings demonstrates the scalability of our approach. Yet, challenges like compatibility concerns and processing overheads underscore the ongoing necessity for system evolution to address changing methodologies.

Transitioning towards an asynchronous, decoupled, and parallel processing architecture might offer an advancement in performance and cost-efficiency.

Our methodology has yielded promising preliminary insights into infrastructure utilisation. The unexpected incorporation of an industrial strike, although unintentional, has provided evidence of a potential methodology. This has emphasised the robustness of our approach, underscoring the system's adaptability to real-world dynamics. Comparing the results to idealised models and simulations could grant clarity, fostering further development and validation.

Reflecting on the project, marked progress has been made in comprehending railway infrastructure utilisation on the UK infrastructure and delineating the divergence between planned and actual train paths. While geospatial methods were crucial in this endeavour, data quality and accessibility dictated the precision of our analysis.

A key challenge, and also an objective, was the system's validation against real-time and empirical railway operations. Detected discrepancies, while indicative of the system's efficacy, underline the need for perpetual validation and continuous improvement of these systems.

Objectively, a geospatial infrastructure was established, merging geographical and logical network data. This crucial step enabled the calculation of a train service's geospatial path and temporal network forecasts. Evaluating this infrastructure's utilisation yielded insights into train service paths, revealing some discrepancies in utilisation.

In conclusion, the project has largely fulfilled its goals. Encountered challenges and unforeseen events have doubled as learning opportunities, paving the way for refinement. The laid groundwork is solid, with multiple prospects for future enhancements to meet the dynamic needs of the railway industry. A vision of an efficient and optimised railway network is attainable, but grasping its current performance and latent potential is crucial. As the railway domain evolves, our systems and methodologies must adapt, ensuring a sustainable and appealing mode of public transport.

8 Future Work

Further work is certainly needed to build on the solid foundations in this project to improve the utilisation calculation and identify if any further insights can be gleamed from improved outputs. A future focus should prioritise both data availability and the quality of source and derived datasets. By developing methods to cleanse and identify data errors iteratively, we can substantially enhance the quality of utilisation outputs.

Improvements in compiling the geospatial model are crucial. This includes the use of advanced methodologies to establish relationships among data layers, completing the desired datasets. Furthermore, methods that bolster dataset quality and enable self-maintenance would be invaluable. Such advancements would support automated, continual improvements, aligning with the dynamic nature of the railway network and reducing manual data preparation, leading to multiple benefits.

Moreover, the geospatial infrastructure model offers areas for refinement. Incorporating additional features – like maximum line speed, gradient, and other pertinent characteristics – would bolster the calculation of investigated capacity metrics and align with current research trends. These enhancements would be pivotal for future modelling and simulation tasks that depend on these features.

While the prediction model has largely succeeded in forecasting the path through the network, there's room for enhancement. This could be achieved by integrating more conventional graph network techniques, and by combining the probabilistic model with other models.

One particularly promising avenue is the real-time application of the prediction model. The modal next step and mean duration have provided a robust foundation. Expanding on the predicted duration, and comparing the confidence intervals of a probabilistic model for each step duration, could usher in an almost real-time monitoring and incident detection system. This system would be sensitive to the minutest of train service behaviour changes, detecting deviations almost instantly and at greater resolution of current systems through the use of statistical calculations. Such rapid detection capabilities could revolutionise railway operations management. Using clustering methods, possibly through spatial analysis, identification and categorisation of delays or events would be possible, streamlining what is currently a manual inspection and monitoring process susceptible to human distractions as workload on operators in the operations community increases.

Further improvements can involve refining the prediction models for the next steps by using mean values as features. This would lead to a precise prediction model with a short time horizon, functioning in real-time to forecast forthcoming steps while considering nearby trains as opposed to solely relying on a train's planned timetable.

Lastly, the geospatial model developed in this project has broader applications, especially in real-time modelling and simulation. The network's spatial nature allows for microscopic modelling that can be validated against macroscopic empirical data. Such validation not only reaffirms the models but could also be integrated into a digital twin of the train service. This integration offers a trifecta: monitoring, prediction, and optimisation based on modelling and simulating either a current or historical system state.

References

£1 billion technology investment to bring railway into 21st century (2022). Press Release. Department for Transport. URL: https://www.gov.uk/government/news/1-billion-technology-investm%20ent-to-bring-railway-into-21st-century.

Abril, M. et al. (2008). "An assessment of railway capacity". In: *Transportation Research Part E: Logistics and Transportation Review* 44.5, pp. 774–806.

Azadi Moghaddam Arani, A., Jolai, F., and Nasiri, M. M. (2019). "A multi-commodity network flow model for railway capacity optimization in case of line blockage". In: *International Journal of Rail Transportation* 7.4, pp. 297–320.

Balfe, N. et al. (2011). "Applying human factors in the design of future rail systems". In: *Human Centred Automation*, pp. 239–250.

Boysen, H. (2013). "General model of railway transportation capacity". In: *13th International Conference on Design and Operation in Railway Engineering (Comprail)*, pp. 335–347.

Burdett, R. L. (2015). "Multi-objective models and techniques for analysing the absolute capacity of railway networks". In: *European Journal of Operational Research* 245.2, pp. 489–505.

Burdett, R. L. and Kozan, E. (2006). "Techniques for absolute capacity determination in railways". In: *Transportation Research Part B: Methodological* 40.8, pp. 616–632.

Cacchiani, V. et al. (2014). "Operations Research in passenger railway transportation". In: *European Journal of Operational Research* 235.3, pp. 591–608.

Corman, F. and Henken, J. (2022). "Estimating aggregate railway performance from realized empirical data: Literature review, a test case and a research roadmap". In: *Journal of Rail Transport Planning & Management* 22, p. 100316.

Department for Transport (2018). *Rail Factsheet - Rail Statistics Overview: 2019*. UK Government. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment%5C%5Fdata/file/851082/rail-factsheet-2019.pdf.

— (2022). *Railways Act 2005 statement: High Level Output Specification*. URL: https://assets.publishing.service.gov.uk/government/uploads/sys%20tem/uploads/attachment%5C%5Fdata/file/1121207/railways-act-2005-statement-high-level-output-specification-2022-web-version.pdf.

Dijkstra, E. W. (1959). "A note on two problems in connexion with graphs". In: *Numerische mathematik* 1.1, pp. 269–271.

Dingler, M. H., Lai, Y.-C., and Barkan, C. P. (2009). "Impact of train type heterogeneity on single-track railway capacity". In: *Transportation Research Record* 2117.1, pp. 41–49.

Francesco, R., Gabriele, M., and Stefano, R. (2016). "Complex railway systems: capacity and utilisation of interconnected networks". In: *European transport research review* 8.4, pp. 1–21.

Gavspark, J., Abramovic, B., and Hals, M. (2015). "New graphical approach to railway infrastructure capacity analysis". In: *Promet-Traffic&Transportation* 27.4, pp. 283–290.

Goverde, R. M., Corman, F., and D'Ariano, A. (2013). "Railway line capacity consumption of different railway signalling systems under scheduled and disturbed conditions". In: *Journal of rail transport planning & management* 3.3, pp. 78–94.

Greater London Authority (2020). *Crossrail - Project at a Glance*. Report. Greater London Authority. URL: https://www.london.gov.uk/sites/default/files/crossrail%5C%5F-%5C%5%20Fproject%5C%5Fat%5C%5Fa%5C%5Fglance.pdf.

Guo, B. et al. (2016). "A study on the practical carrying capacity of large high-speed railway stations considering train set utilization". In: *Mathematical Problems in Engineering* 2016.

Harrod, S. (2009). "Capacity factors of a mixed speed railway network". In: *Transportation Research Part E: Logistics and Transportation Review* 45.5, pp. 830–841.

HS2 (2023). *HS2 Funding*. Web Page. HS2. URL: https://www.hs2.org.uk/what-is-hs2/hs2-funding/.

Jamili, A. (2018). "Computation of practical capacity in single-track railway lines based on computing the minimum buffer times". In: *Journal of Rail Transport Planning & Management* 8.2, pp. 91–102.

Jensen, L. W. (2015). "An optimisation framework for determination of capacity in railway networks". In: *Proceedings of the 13th Conference on Advanced Systems in Public Transport (caspt2015), Rotterdam, Netherlands*.

Jensen, L. W., Landex, A., et al. (2017). "Strategic assessment of capacity consumption in railway networks: Framework and model". In: *Transportation Research Part C: Emerging Technologies* 74, pp. 126–149.

Jensen, L. W., Schmidt, M., and Nielsen, O. A. (2020). "Determination of infrastructure capacity in railway networks without the need for a fixed timetable". In: *Transportation Research Part C: Emerging Technologies* 119, p. 102751.

Jia, F. et al. (2021). "Passenger-oriented subway network capacity calculation and analysis based on simulation". In: *Transportation Letters* 13.8, pp. 555–567.

Krehl, C. and Balfe, N. (2014). "Cognitive workload analysis in rail signalling environments". In: *Cognition, technology & work* 16, pp. 359–371.

Lai, Y. C. and Lin, Y. J. (2016). "Using profit-maximizing capacity framework and models for railway capacity management". In: *Journal of Transportation Engineering* 142.10, p. 04016045.

Lai, Y.-C., Huang, Y.-A., and Chu, H.-Y. (2014). "Estimation of rail capacity using regression and neural network". In: *Neural Computing and Applications* 25, pp. 2067–2077.

Lai, Y.-C., Liu, Y.-H., and Lin, T.-Y. (2012). "Development of base train equivalents to standardize trains for capacity analysis". In: *Transportation Research Record* 2289.1, pp. 119–125.

Landex, A. (2009). "Evaluation of railway networks with single track operation using the UIC 406 capacity method". In: *Networks and Spatial Economics* 9, pp. 7–23.

Li, Y. et al. (2019). "Tactical capacity assessment of a high-speed railway corridor with high heterogeneity". In: *Proc., 8th Int. Conf. on Railway Operations Modelling and Analysis. Delft, Netherlands: International Association of Railway Operations Research.*

Liao, Z. et al. (2021). “Railway capacity estimation considering vehicle circulation: Integrated timetable and vehicles scheduling on hybrid time-space networks”. In: *Transportation Research Part C: Emerging Technologies* 124, p. 102961.

Lin, T.-Y., Lin, Y.-C., and Lai, Y.-C. “. (2020). “Base train equivalents for multiple train types based on delay-based capacity analysis”. In: *Journal of Transportation Engineering, Part A: Systems* 146.5, p. 04020023.

Ljubaj, I. and Mlinarić, T. J. (2019). “The possibility of utilising maximum capacity of the double-track railway by using innovative traffic organisation”. In: *Transportation Research Procedia* 40, pp. 346–353.

Lusby, R. M. et al. (2011). “Railway track allocation: models and methods”. In: *OR spectrum* 33.4, pp. 843–883.

Mikulčić, M. and Mlinarić, T. J. (2021). “Railway Capacity Enhancement with Modern Signalling Systems–A Literature Review”. In: *Promet-Traffic&Transportation* 33.1, pp. 141–152.

Murali, P. et al. (2010). “A delay estimation technique for single and double-track railroads”. In: *Transportation Research Part E: Logistics and Transportation Review* 46.4, pp. 483–495.

Mussone, L. and Calvo, R. W. (2013). “An analytical approach to calculate the capacity of a railway system”. In: *European Journal of Operational Research* 228.1, pp. 11–23.

Nachtigall, K. (1995). “Time depending shortest-path problems with applications to railway networks”. In: *European Journal of Operational Research* 83.1, pp. 154–166.

Network Rail (2014). *ECML 2020 Capacity - Timetable Assessment*. Network Rail. URL: https://www.orr.gov.uk/sites/default/files/om/ecml-2020-capacity-timetable-assessment-dec-2014.pdf.

— (2017). *How rail timetabling works*. Network Rail. URL: https://www.networkrail.co.uk/wp-content/uploads/2017/11/How-rail-timetabling-works-factsheet.pdf.

— (2021a). *Castlefield Corridor Congested Infrastructure: Capacity Enhancement Plan*. Network Rail. URL: https://www.networkrail.co.uk/wp-content/uploads/2021/02/Castlefield-Corridor-Congested-Infrastructure-Capacity-Enhancement-Plan-26-February-2021.pdf.

— (2021b). *Weather Resilience and Climate Change Adaptation Report*. Network Rail. URL: https://safety.networkrail.co.uk/wp-content/uploads/2023/02/Climate-Change-Adaptation-Report-2021.pdf.

— (2023a). *London North Eastern Timetable Planning Rules 2023*. Network Rail. URL: https://sacuksprodnrdigital0001.blob.core.windows.net/operational-rules/Operational%20Rules/TPR%202023/V4/TPR%202023%20V4%20LNE.pdf.

— (2023b). *Timetable Planning Rules 2023*. Network Rail. URL: https://www.networkrail.co.uk/industry-and-commercial/information-for-operators/operational-rules/.

Office of Rail and Road (2014). *Technology Adoption Case Studies - Targeted Assurance Review*. Office of Rail and Road. URL: https://www.orr.gov.uk/media/23300/download.

Office of Rail and Road (2020). *2020 Annual Report on the Railways in the UK*. Office of Rail and Road. URL: https://www.orr.gov.uk/sites/default/files/2021-06/orr-annual-report-and-accounts-2020-21-web.pdf.

Pachl, J. (2002). *Railway Operation and Control*. VTD Rail Pub. URL: https://books.google.co.uk/books?id=flrbAAAAMAAJ.

Pickup, L., Wilson, J., and Lowe, E. (2010). "The operational demand evaluation checklist (ODEC) of workload for railway signalling". In: *Applied ergonomics* 41.3, pp. 393–402.

Pouryousef, H. and Lautala, P. (2015). "Hybrid simulation approach for improving railway capacity and train schedules". In: *Journal of Rail Transport Planning & Management* 5.4, pp. 211–224.

Railway Delivery Group (2023). *Delay Attribution Review*. Railway Delivery Group. URL: https://www.orr.gov.uk/sites/default/files/2020-09/rdg-delay-attribution-review-report-2020-09-28.pdf.

Rais, N. M. et al. (2020). "Measuring railway track capacity and utilization for KERETAPI tanah melayu berhad (ktmb)". In: *Advances in Transportation and Logistics Research* 3, pp. 908–916.

Riejos, F. A. O. et al. (2016). "Analyzing the theoretical capacity of railway networks with a radial-backbone topology". In: *Transportation Research Part A: Policy and Practice* 84, pp. 83–92.

Sameni, M. K., Dingler, M., et al. (2011). "Profit-generating capacity for a freight railroad". In: *Transportation Research Board 90th Annual Meeting*.

Sameni, M. K., Landex, A., and Preston, J. (2011). "Developing the UIC 406 method for capacity analysis". In: *4th international seminar on railway operations research*. Italy Rome.

Sameni, M. K. and Moradi, A. (2022). "Railway capacity: A review of analysis methods". In: *Journal of Rail Transport Planning & Management* 24, p. 100357.

System Operator (2019). *Data Architecture Reference Model*. Route Services: Network Rail. URL: https://www.networkrail.co.uk/wp-content/uploads/2019/12/Data-Architecture-Reference-Model.pdf.

The UK's Geospatial Strategy, 2020 to 2025 (2020). Report. Geospatial Commission. URL: https://assets.publishing.service.gov.uk/government/uploads/sy%20stem/uploads/attachment%5C%5Fdata/file/894755/Geospatial%5C%5FStrategy.pdf.

Transport Committee (2016). *The future of rail: Improving the rail passenger experience*. Journal Article. House of Commons. URL: https://publications.parliament.uk/pa/cm201617/cmselect/cmtra%20ns/64/64.pdf.

UIC Code 406, 1st Edition (2004). Report. International Union of Railways. URL: https://tamannaei.iut.ac.ir/sites/tamannaei.iut.ac.ir/files/fi%20les%5C%5Fcourse/uic406%5C%5F2004.pdf.

UIC Code 406, 2nd Edition (2013). Report. International Union of Railways. URL: https://tamannaei.iut.ac.ir/sites/tamannaei.iut.ac.ir/files/fi%20les%5C%5Fcourse/uic406%5C%5F2013.pdf.

Valentinovic, L. and Sivilevicius, H. (2014). "Railway line capacity methods analysis and their application in Lithuanian Railways justification". In: *Environmental Engi-*

neering. Proceedings of the International Conference on Environmental Engineering. ICEE. Vol. 9. Vilnius Gediminas Technical University, Department of Construction Economics . . ., p. 1.

Vaswani, A. et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems*, pp. 5998–6008.

Wang, R., Nie, L., and Tan, Y. (2020). "Evaluating line capacity with an analytical UIC code 406 compression method and blocking time stairway". In: *Energies* 13.7, p. 1853.

Weik, N. et al. (2020). "Extending UIC 406-based capacity analysis–New approaches for railway nodes and network effects". In: *Journal of Rail Transport Planning & Management* 15, p. 100199.

Wong, D. (2009). "The Modifiable Areal Unit Problem (MAUP)". In: *The SAGE Handbook of Spatial Analysis*. London: SAGE, pp. 105–123.

Wu, J. and Zhang, C. (2019). "Research on Calculation Method of Passing Capacity of High Speed Railway Station and Design of Simulation Process". In: *Electrical Engineering and Computer Science (EECS)* 1, pp. 39–44.

Yu, B., Yin, H., and Zhu, Z. (2018). "Spatio-temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI).*

Zhong, Q. et al. (2019). "Railway Infrastructure Capacity Utilization Description through Data Integration in Blocking Time Theory". In.

Appendices

A Dataset Summary

Table A.1: Datasets used in analysis with their description, sources and various characteristics

Dataset	Name	Description	Type	Licence	Source	Data Format	Volatility	Periodicity	Daily Volume
Network Infrastructure Model	NIM	A geospatial representation of the track centreline of all network rail tracks	Provided	OGL	FOI	OSGB36 Shape File	Static		50 MB
SMART Database	SMART	A database linking TD Berths to Locations for use in the TRUST Movement messages to determine progress of trains against timetable	Provided	OGL	NROD	Excel Spreadsheet	Dynamic	Biannually	6.3 MB
Signal Assets	Signals	A extract of Network Rails Ellipse Asset Database for Signal assets	Provided	OGL	FOI	Excel Spreadsheet	Static		8.5 MB
NR National Timetable	Timetable	A daily export of the current planned timetable (timetable for next 6 months)	Provided	OGL	NROD	JSON	Dynamic	Daily	1.8 GB
List of Timetable Locations	Locations	A list of timetable locations currently referenced in the timetable	Provided	OGL	Timetable	JSON	Dynamic	Weekly	8 MB
Agreed Timetable	Daily Timetable	A list of the timetabled serviced due to run on a certain day	Derived	Derived OGL	Timetable	JSON	Dynamic	Daily	400 MB
Current Plan	Plan	An up to date plan for the days train serviced (considering cancellations and VSTPs)	Derived	Derived OGL	Agreed Timetable, TRUST Feed, VSTP	JSON	Dynamic	Minute	1.7 GB
Train Describer Feed	TD	Signal berth data feed from the Train Describers showing raw train movements between signals	Provided	OGL	NROD	JSON	Data Feed	Second	500 MB
TRUST Feed	TRUST	Messages from the TRUST system reporting states of a train services lifecycle	Provided	OGL	NROD	JSON	Data Feed	Second	500 MB
TRUST Activations	ACT	Messages stating a train service has been actiated, most services are called automatically (passenger services) with some being activated as required	Provided	OGL	TRUST	JSON	Dynamic	Minute	70 MB
TRUST Movements	MVT	Messages recording a trains movement and status against its planned timetable	Provided	OGL	TRUST	JSON	Dynamic	Minute	400 MB
TRUST Calculations	CAN	Messages reporting a cancellation of a train service	Provided	OGL	TRUST	JSON	Dynamic	Minute	30 MB
VSTP Feed	VSTP	Messages from Very Short Term Planning service for train service not in the published timetable	Provided	OGL	NROD	JSON	Dynamic	Minute	30 MB
List of TD Steps	TD Steps	List of unique TD Steps between TD Berths	Derived	Derived OGL	TD	JSON	Static	Quarterly	1 MB
List of TD Berths	TD Berths	List of unique TD Berths	Derived	Derived OGL	TD	JSON	Static	Quarterly	1 MB
Train Describer Network	TD Network	Graph of TD Berths (Nodes) and TD Steps (Edges)	Derived	Derived OGL	TD	NetworkX	Static	Quarterly	4 MB

B Project Data

All static data and scripts used by this project are hosted in a public GitHub repository to share with the wider community inline with UK RI's Open Research culture all data and code is released in accordance with the FAIR Data Principles.

The processed raw data due to its size will not be available but could be reproduced following the provided scripts and reference data provided in the open repository. Some files in the repository have been compressed in Zip or 7z depending on the amount of compression required to accomodate large files into the repository.

The repository is available at Stuart Gordon GitHub Repository - Analysing Railway Infrastructure Utilisation: A Geospatial Approach.

B.1 Project Data Repository Structure

The following document tree details the structure of the repository for guidance:

```
Analysing-Railway-Infrastructure-Utilisation-A-Geospatial-Approach
 ├─ Infrastructure - Files for the cloud infrastructure
 │   ├─ Applications - Applications in the cloud infrastructure
 │   │   ├─ camel - Apache Camel application for retrieving messages from NR
 │   │   ├─ cif_processor - Timetable Processing Script
 │   │   ├─ finished - Realtime train finished routing and pathing process
 │   │   ├─ message_logger - Logging of NR messages for OLAP
 │   │   ├─ message_processor - Processing realtime messages in AWS Lambda
 │   │   ├─ rabbit_mq_config_tool - Configuration tool for Rabbit Broker
 │   │   └─ rabbit_processor - Python function to turn incoming NR messages
 │   │      into Rabbit topics with routing keys
 │   ├─ LambdaLayers - Python libraries for use with AWS Lambda serverless
 │   │  functions
 │   └─ Templates - Contains CloudFormation templates for provisioning cloud
 │      infrastructure
 └─ RealtimeProcessing - Processing files
     └─ lambda - Lambda functions for message processing
```

```
│
│   │
│   ├── mongo_frontend - Front End visualisation tools
│   │   ├── batch_processing - Batch Processing Analysis
│   │   │   ├── routing - Routing Functions
│   │   │   └── utilisation_count - Utilisation Calculation Functions
│   │   ├── berths_import - Processing of Geospatial Data
│   │   └── webapp - Web applications used for visualisations for a variety
│   │       of use cases
│   ├── pika - Python function to turn incoming NR messages into Rabbit topics
│   │   with routing keys
│   ├── rabbit - Rabbit Configuration Tool
│   ├── TDPredictionAndProcessing - TD Prediction Scripts and process
│   ├── timetable - Rabbit Configuration Tool
│   └── train-viewer - Web applications used for visualisations of train
│       data
├── Results - Static result files output from analysis
└── Writeup - Project Submission Files
    ├── Poster - Thesis documentation files
    ├── Presentation - Thesis presentation files
    └── Thesis - Thesis poster files
```

C Normalised Planned, Normalised Actual and Difference Utilisation Maps

C.1 Normalised Planned Utilisation Maps

Normalised Planned Utilisation Maps for the data collection period for each day during the collection period from 11th July 2023 to 23rd July 2023. These are shown in Figures C.1 to C.13.

C.2 Normalised Actual Utilisation Maps

Normalised Actual Utilisation Maps for the data collection period for each day during the collection period from 11th July 2023 to 23rd July 2023. These are shown in Figures C.14 to C.26.

C.3 Difference in Planned to Actual Utilisation Maps

Difference in Planned to Actual Utilisation Maps for the data collection period for each day during the collection period from 11th July 2023 to 23rd July 2023. These are shown in Figures C.27 to C.39.

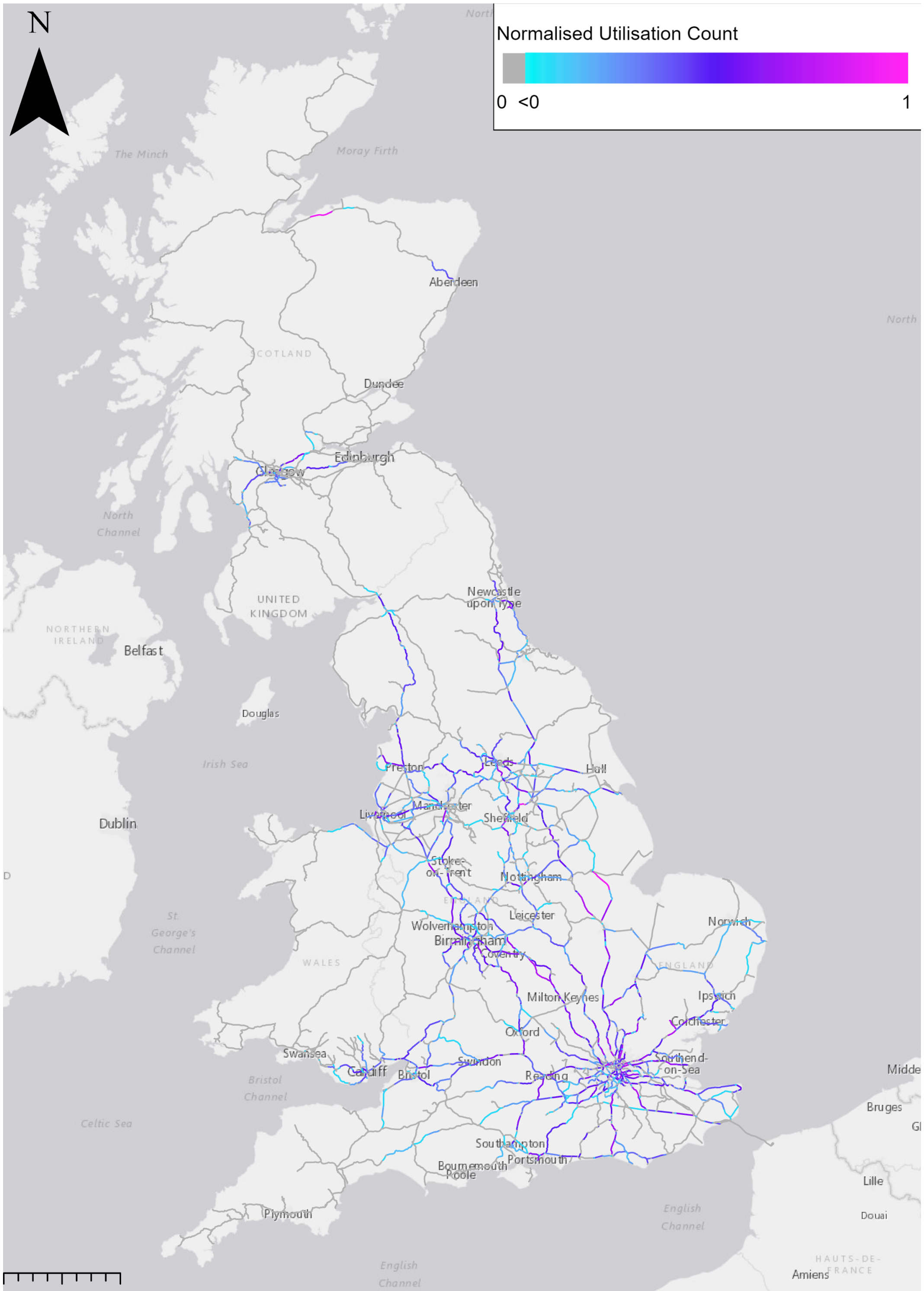


Figure C.1: Normalised Planned Utilisation for 11/07/2023

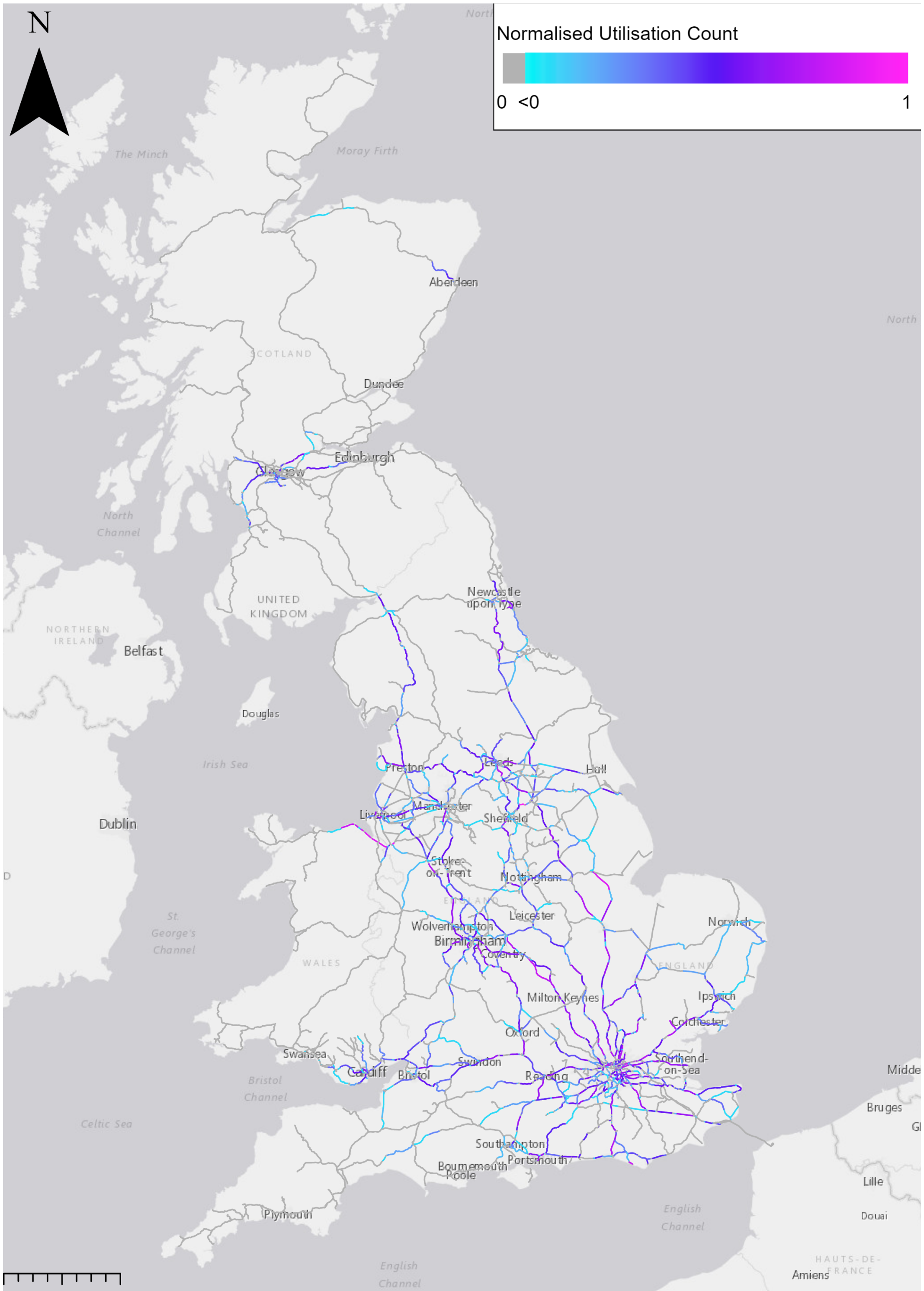


Figure C.2: Normalised Planned Utilisation for 12/07/2023

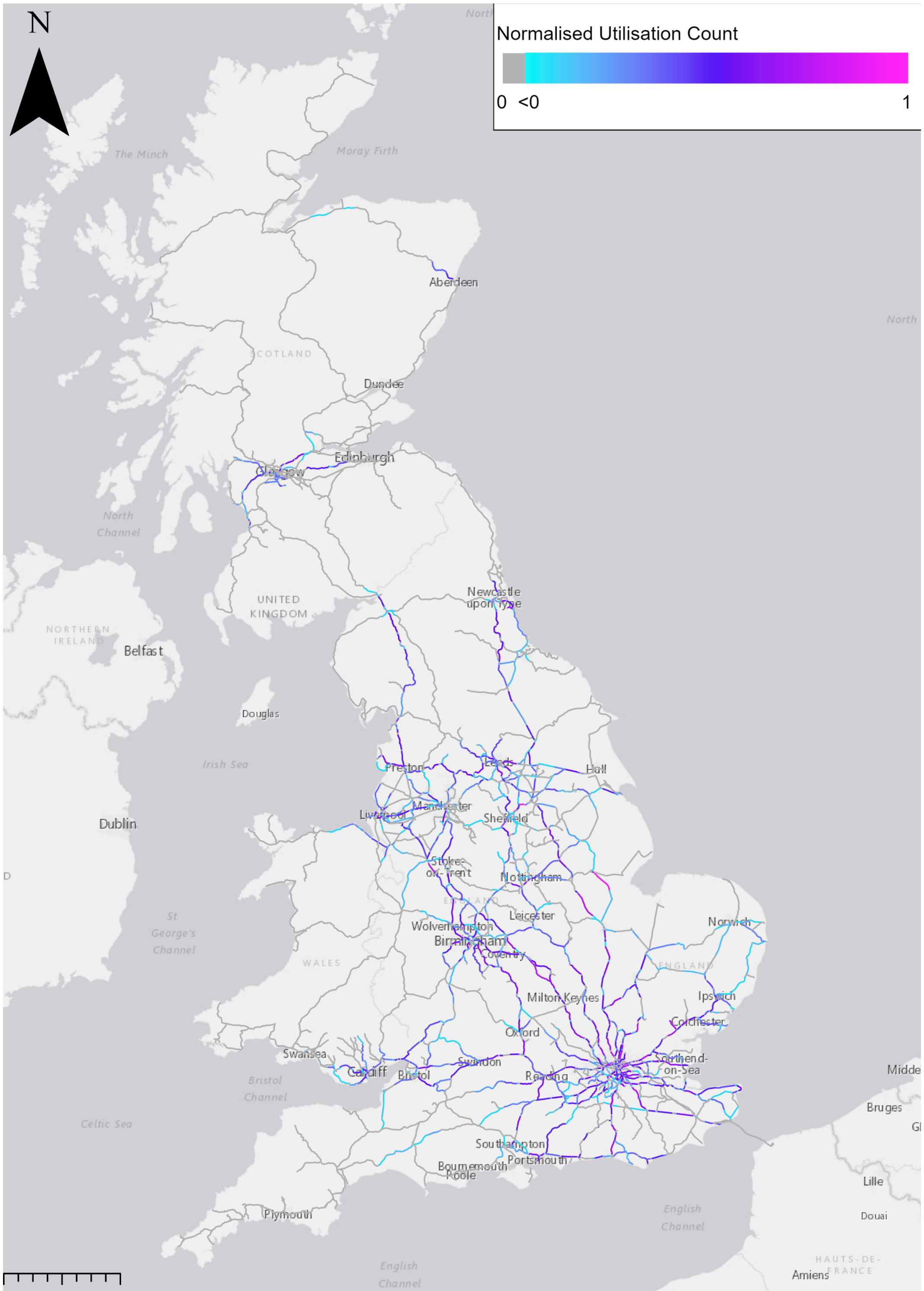


Figure C.3: Normalised Planned Utilisation for 13/07/2023

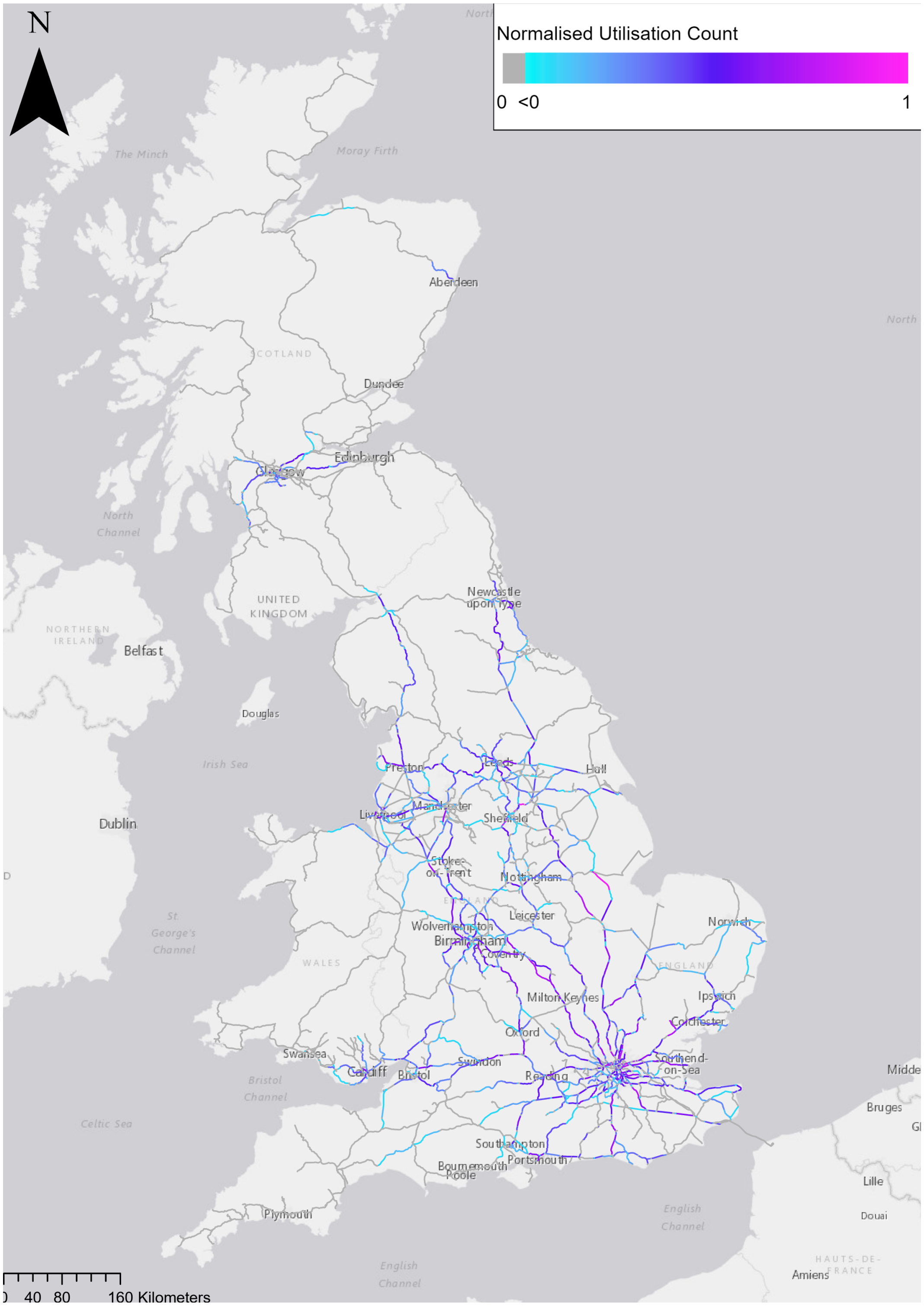


Figure C.4: Normalised Planned Utilisation for 14/07/2023

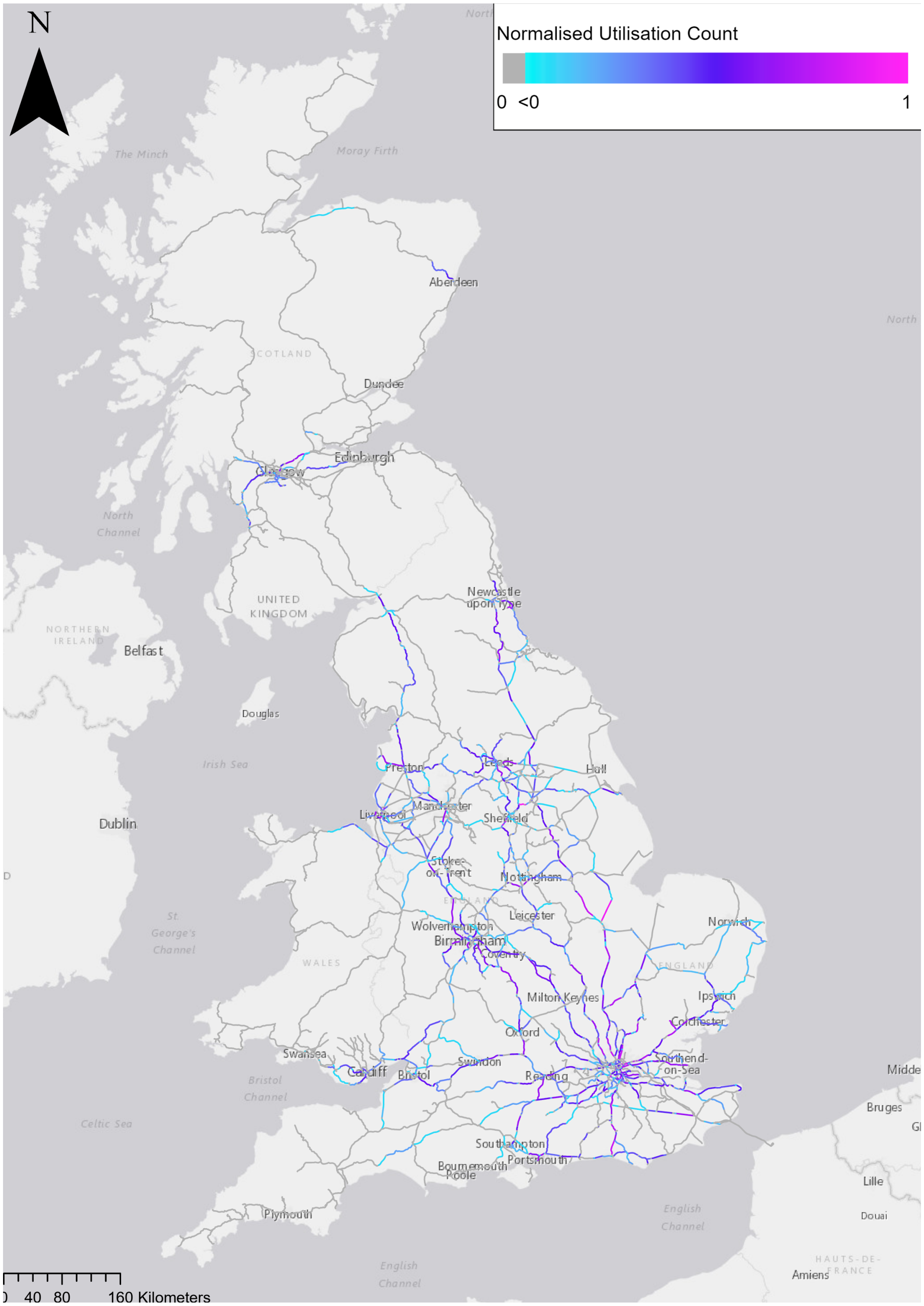


Figure C.5: Normalised Planned Utilisation for 15/07/2023

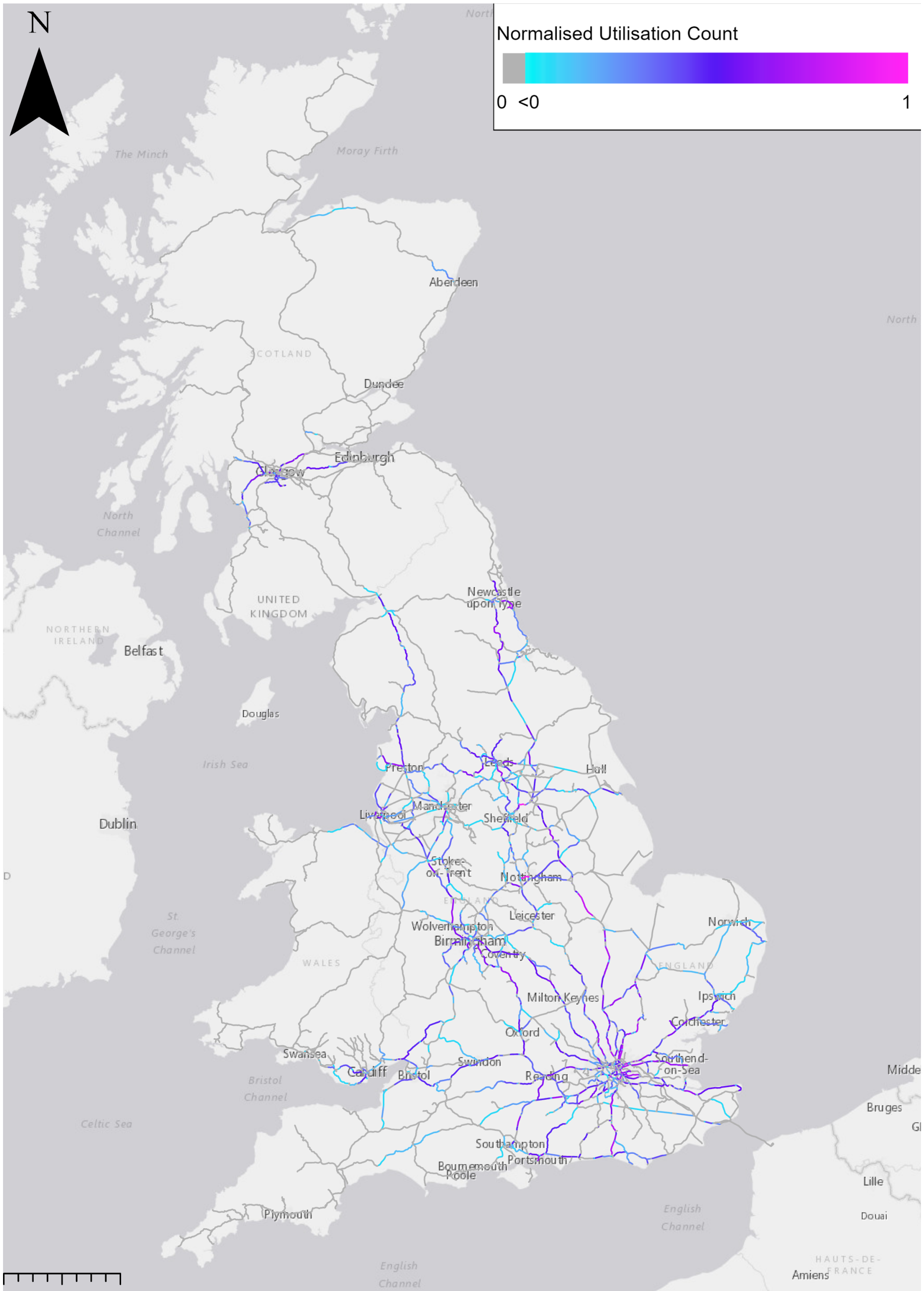


Figure C.6: Normalised Planned Utilisation for 16/07/2023

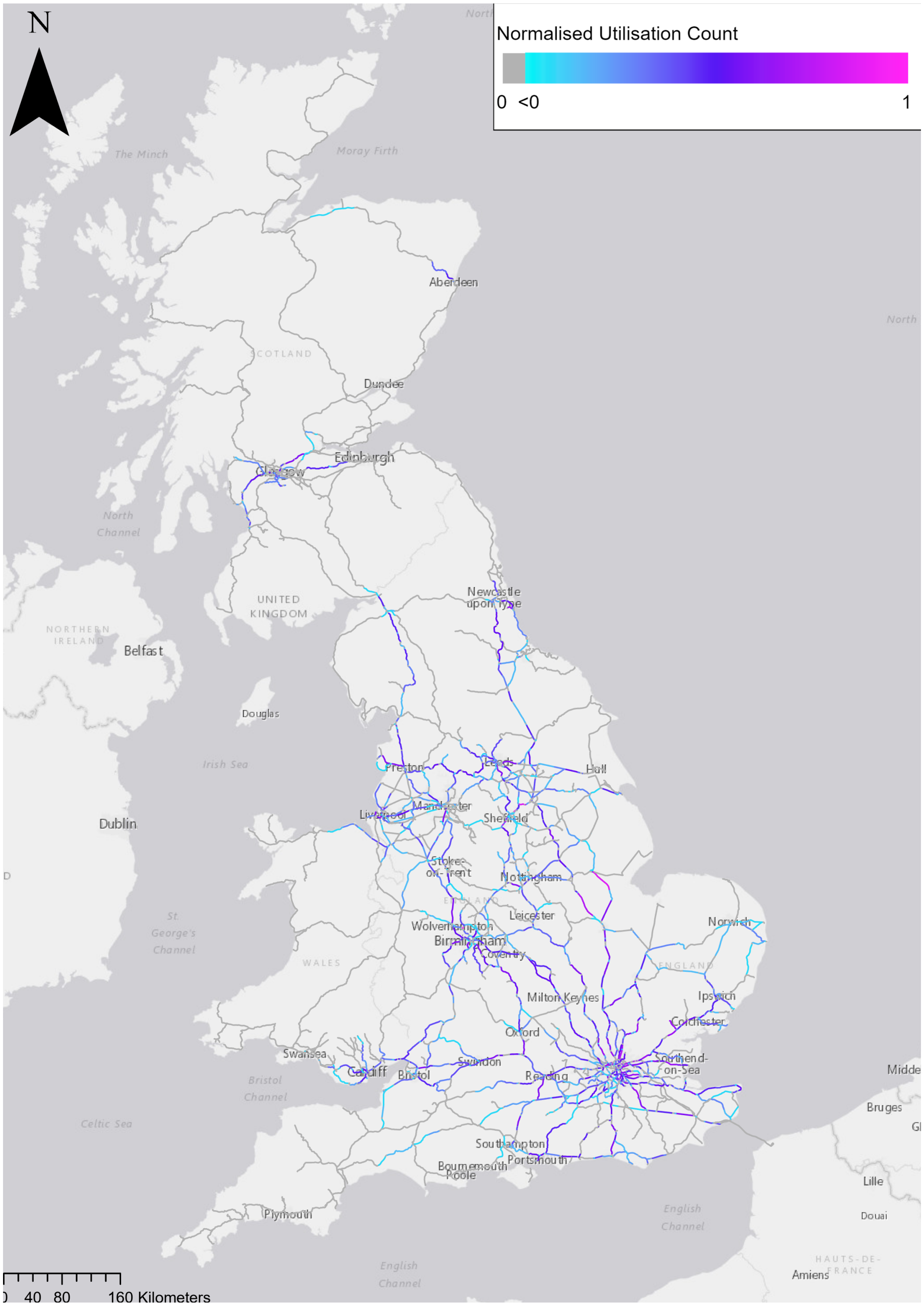


Figure C.7: Normalised Planned Utilisation for 17/07/2023

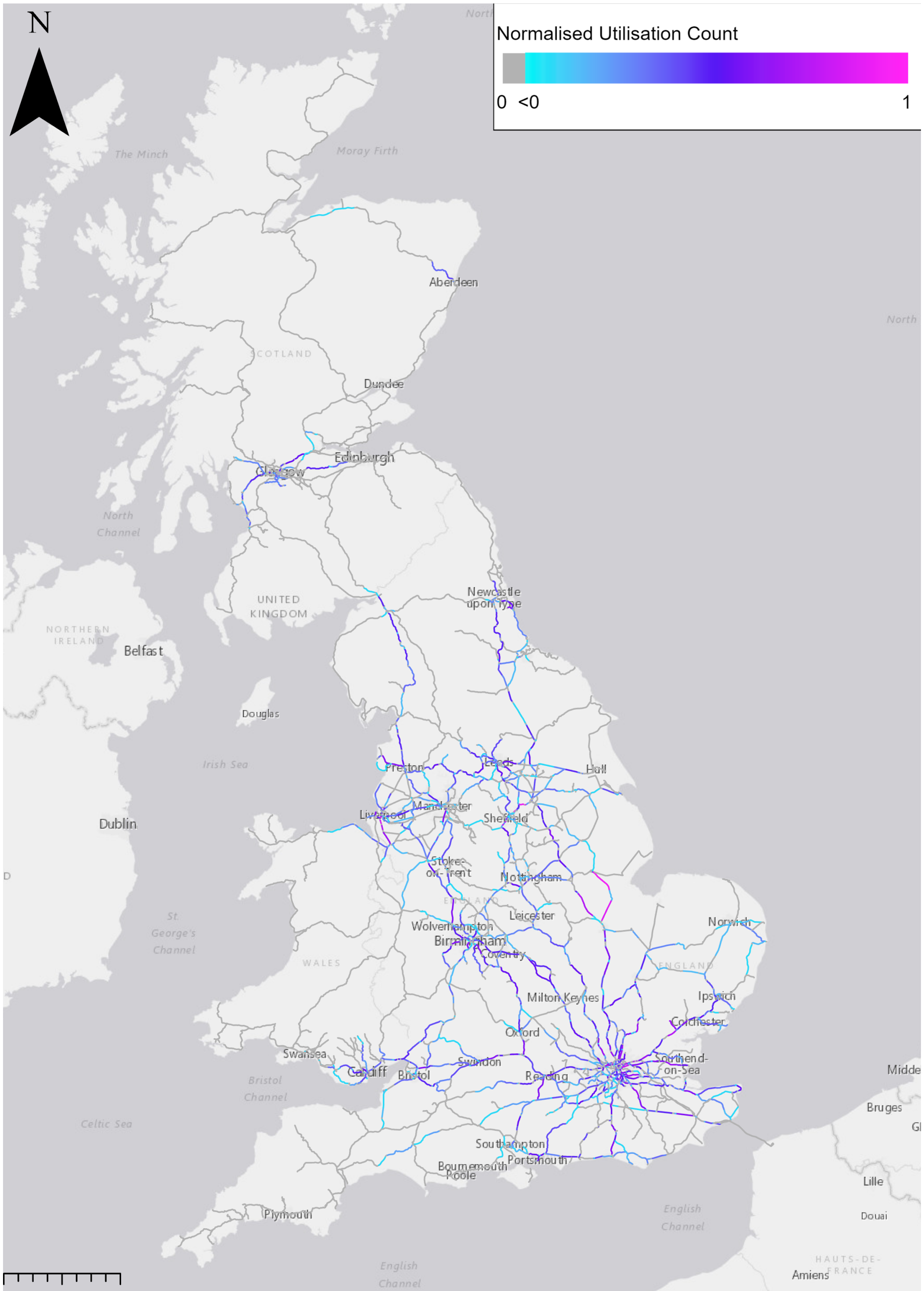


Figure C.8: Normalised Planned Utilisation for 18/07/2023

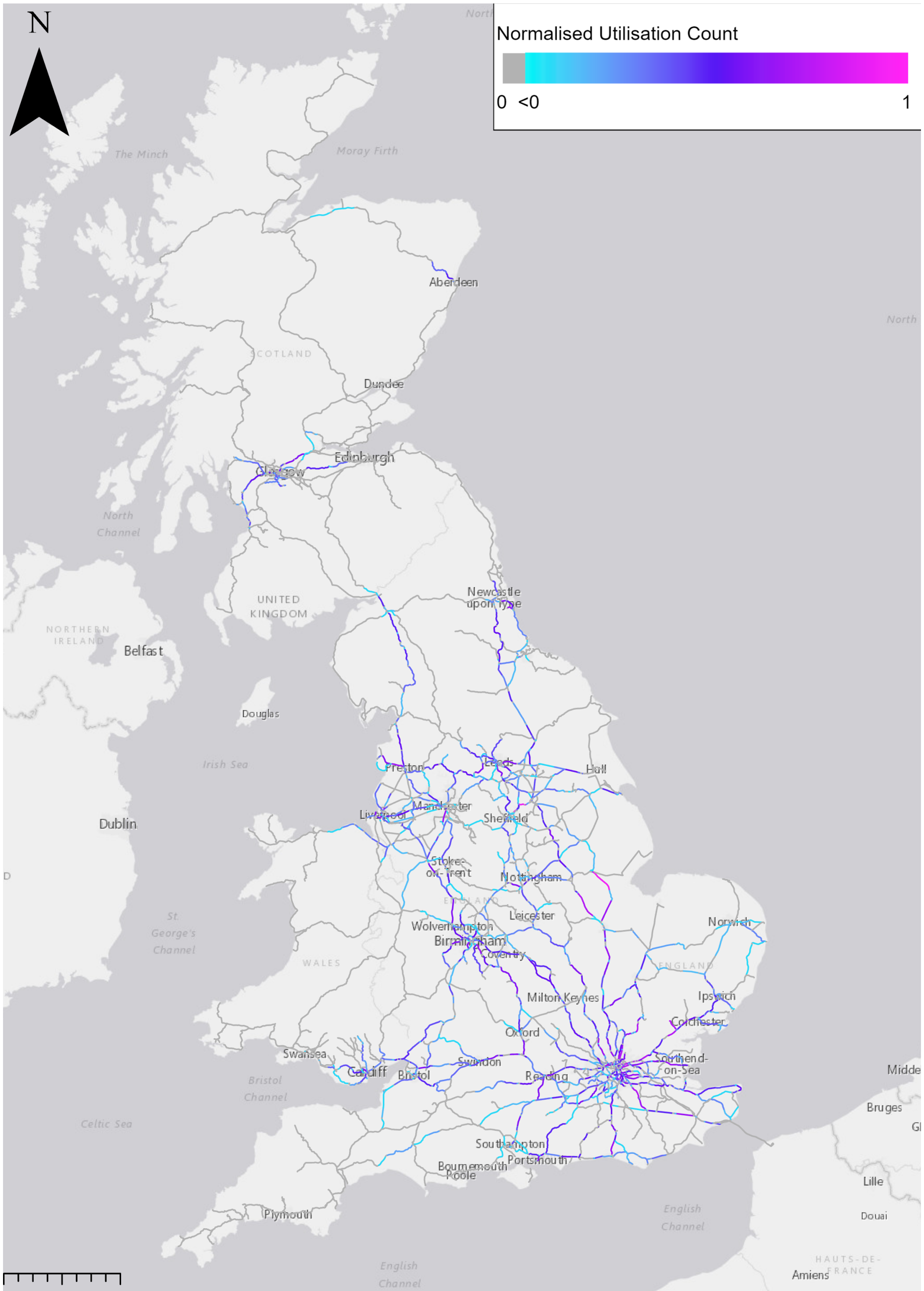


Figure C.9: Normalised Planned Utilisation for 19/07/2023

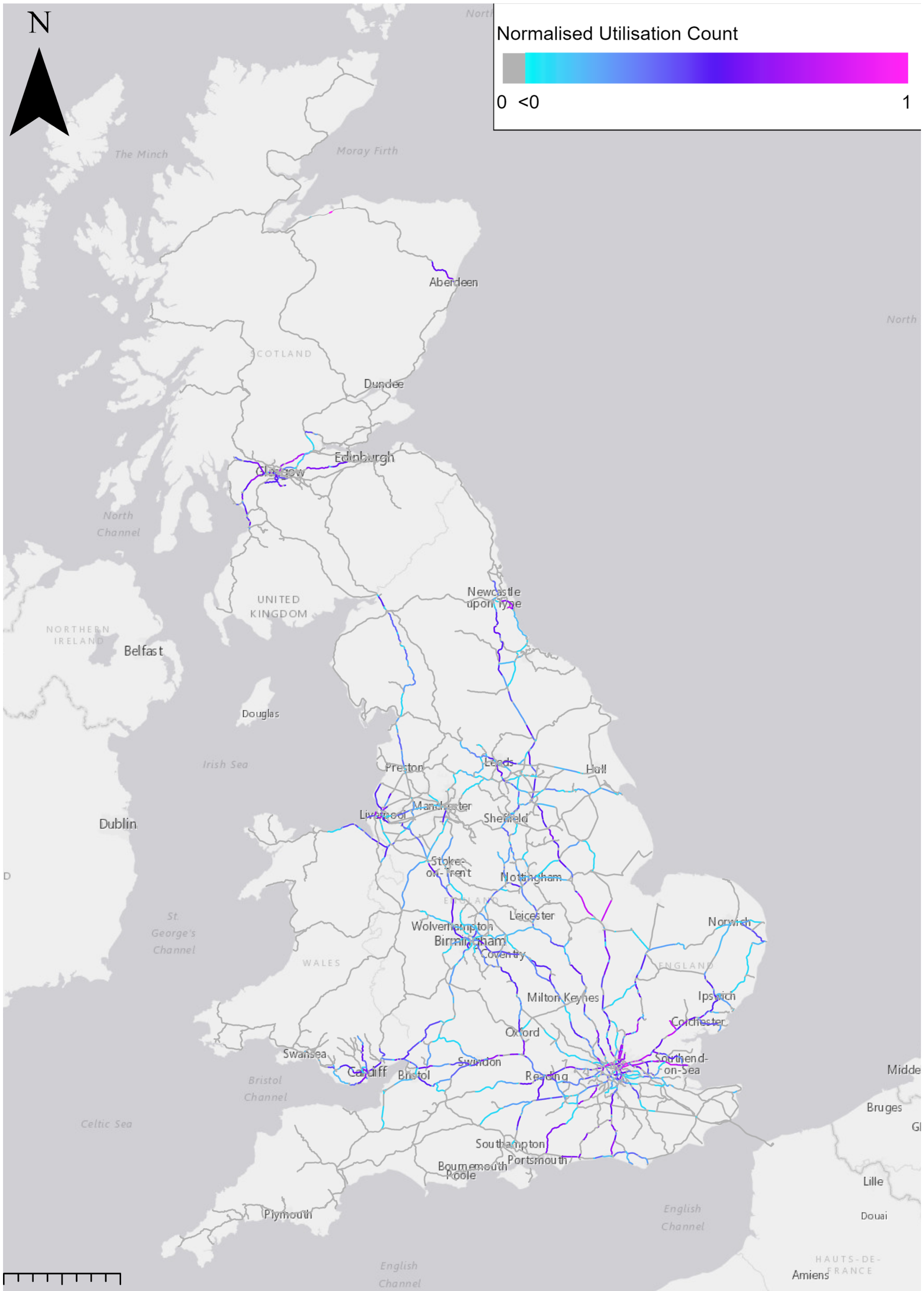


Figure C.10: Normalised Planned Utilisation for 20/07/2023

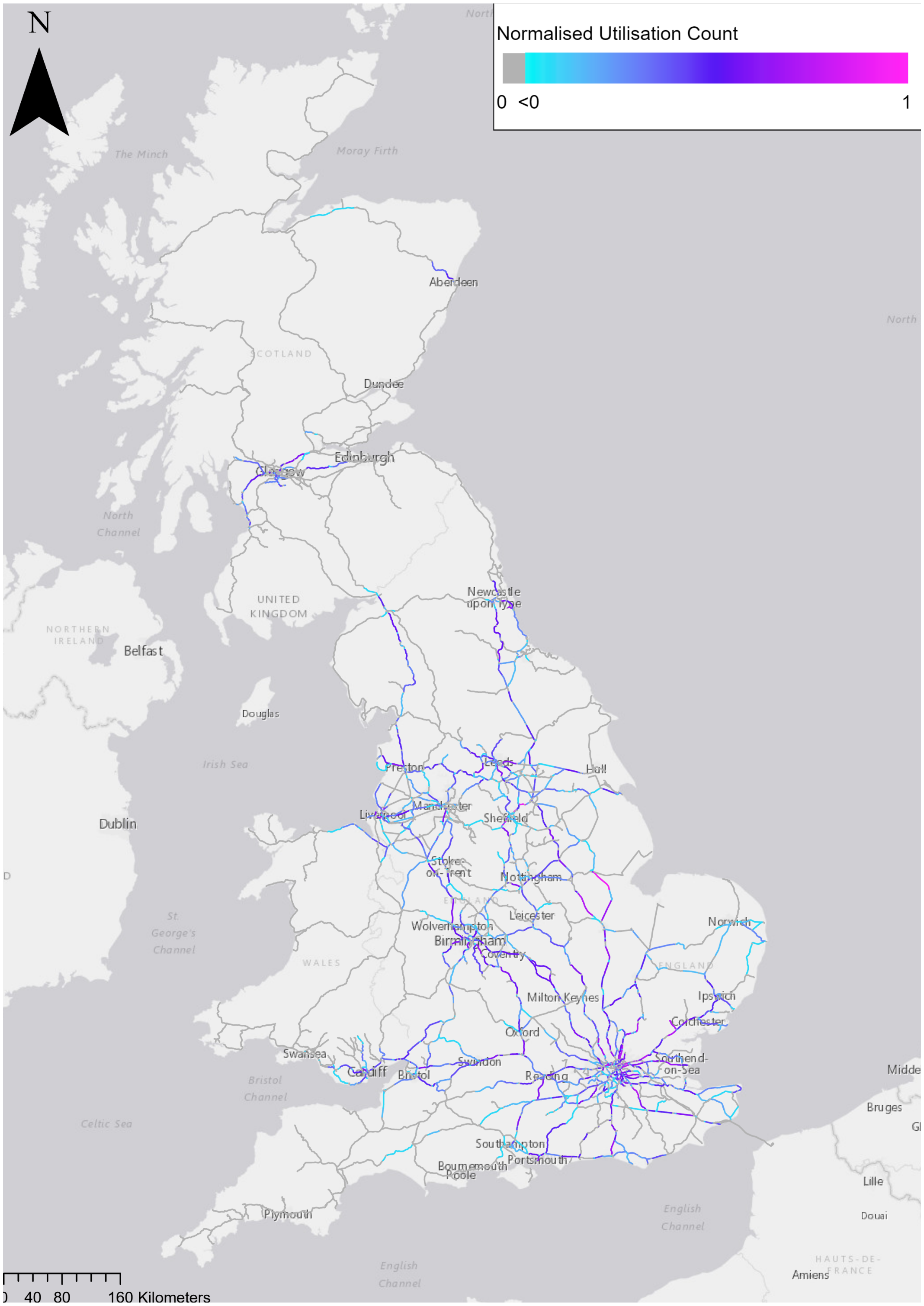


Figure C.11: Normalised Planned Utilisation for 21/07/2023

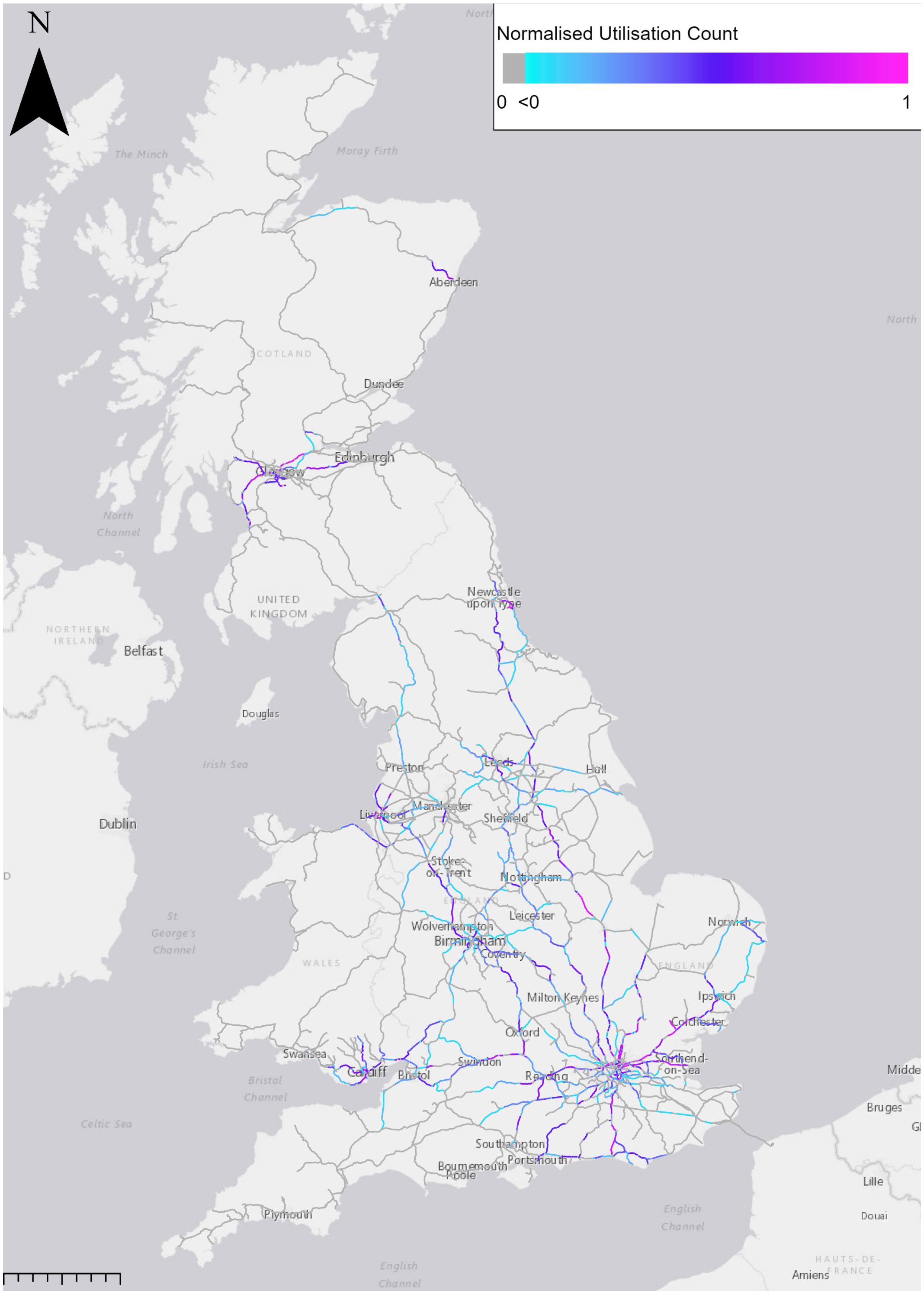


Figure C.12: Normalised Planned Utilisation for 22/07/2023

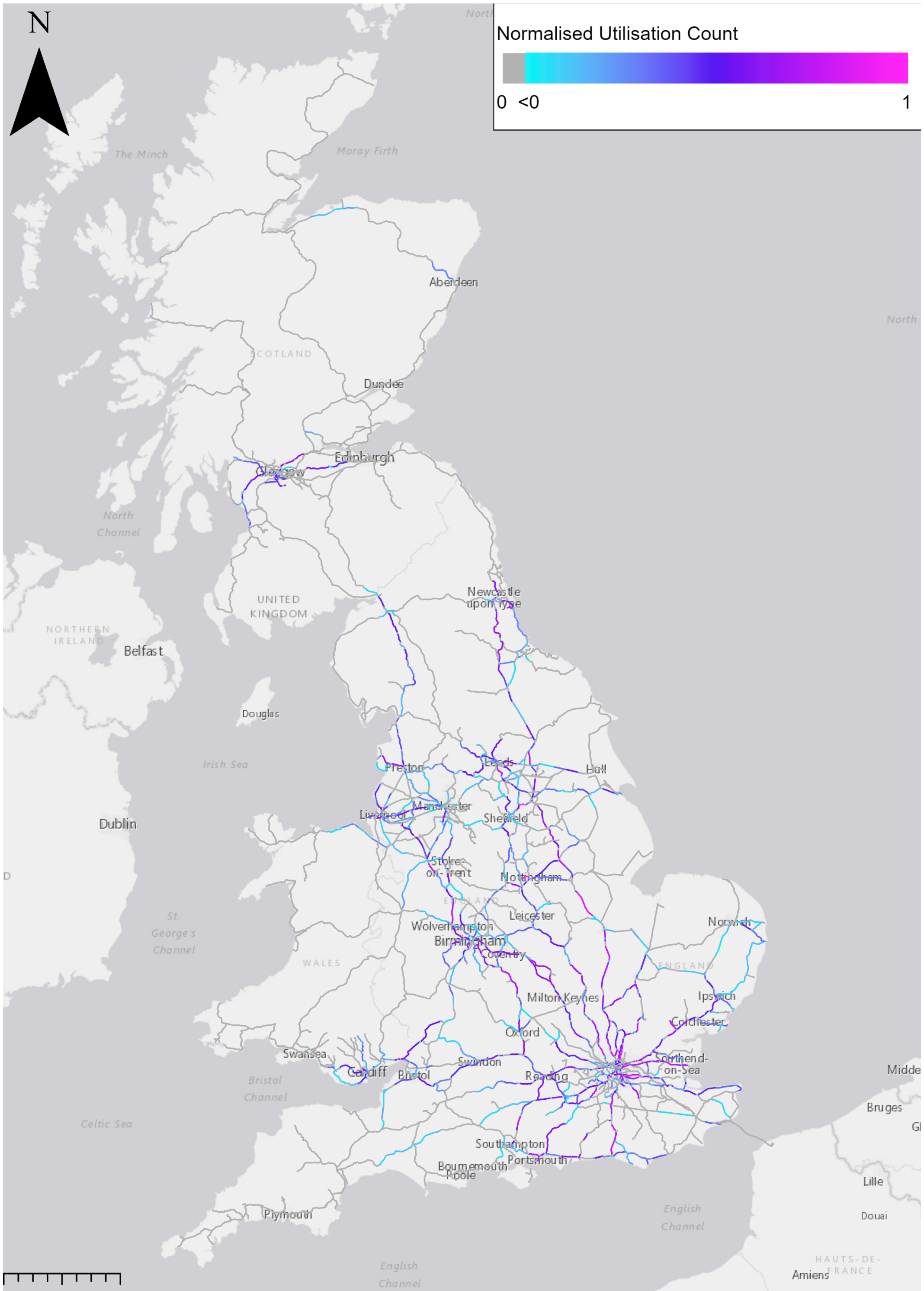


Figure C.13: Normalised Planned Utilisation for 23/07/2023

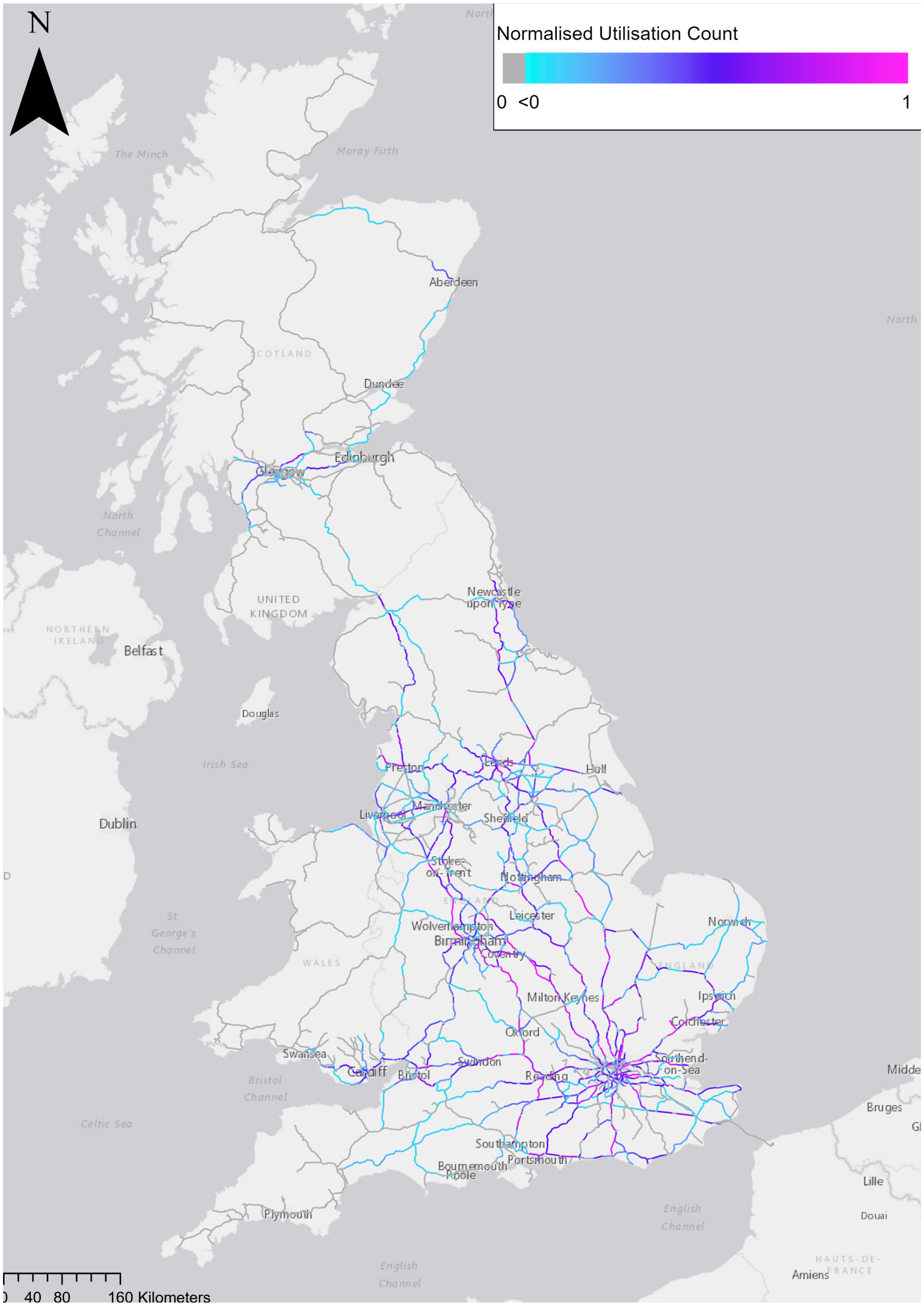


Figure C.14: Normalised Actual Utilisation for 11/07/2023

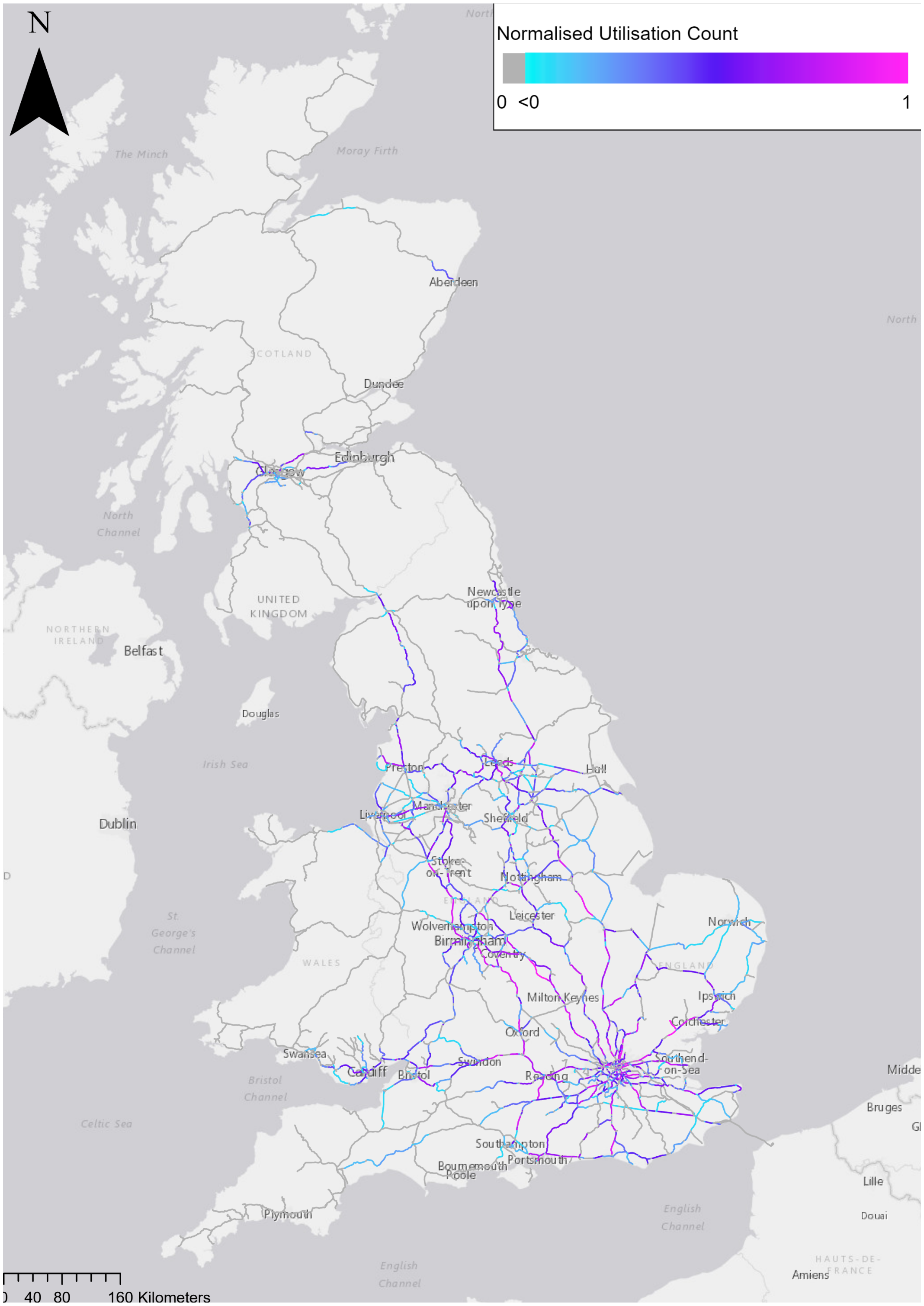


Figure C.15: Normalised Actual Utilisation for 12/07/2023

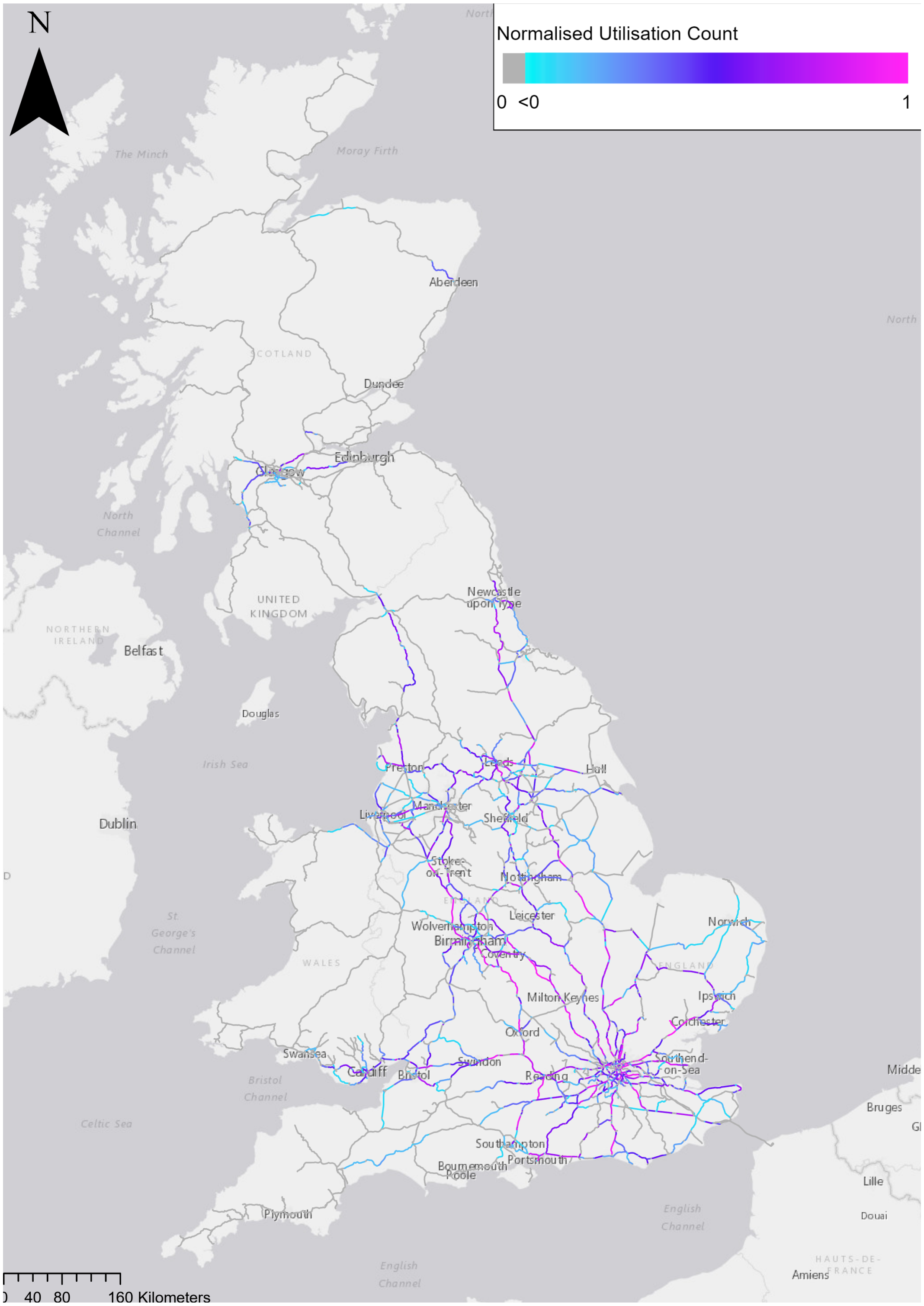


Figure C.16: Normalised Actual Utilisation for 13/07/2023

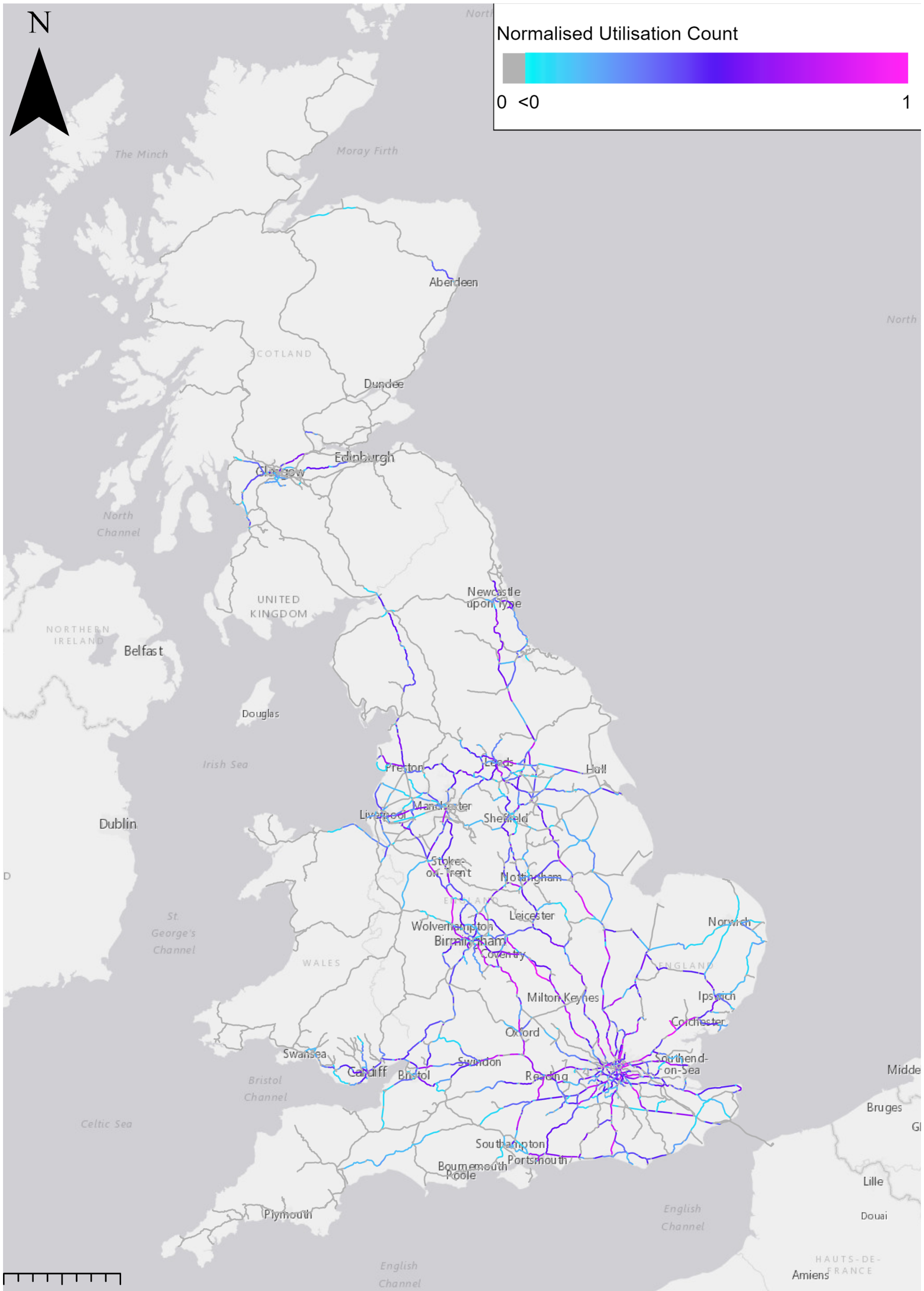


Figure C.17: Normalised Actual Utilisation for 14/07/2023

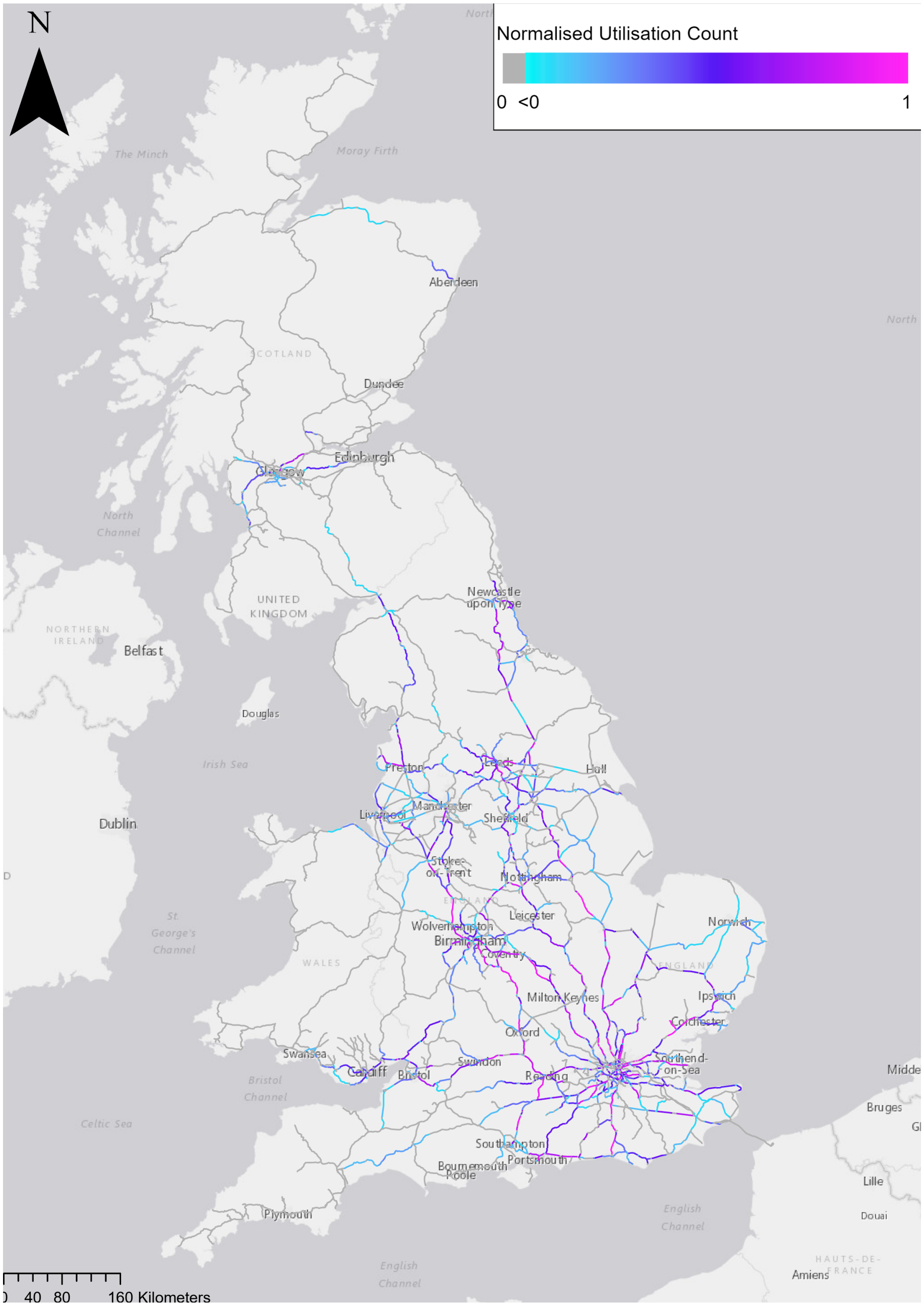


Figure C.18: Normalised Actual Utilisation for 15/07/2023

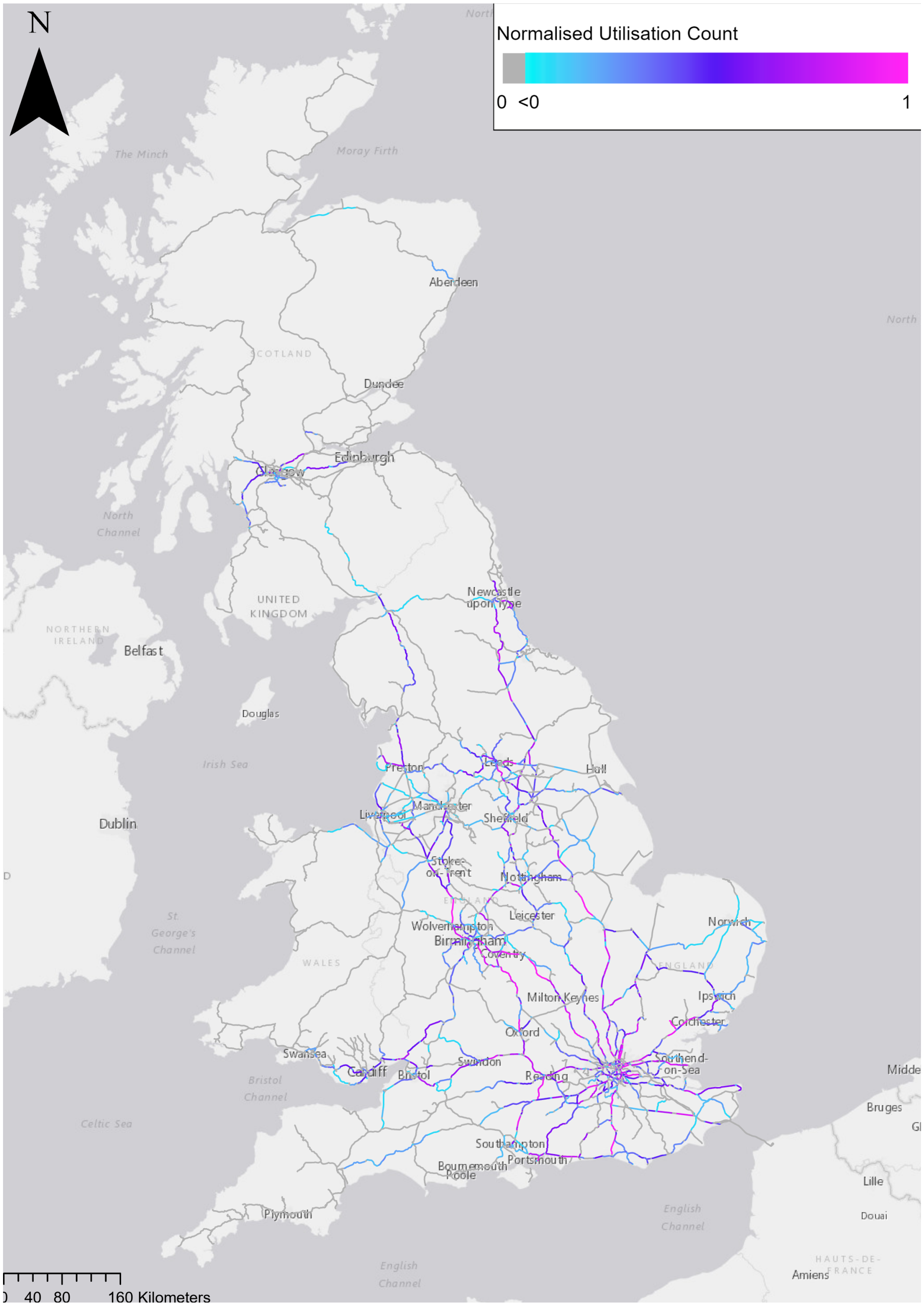


Figure C.19: Normalised Actual Utilisation for 16/07/2023

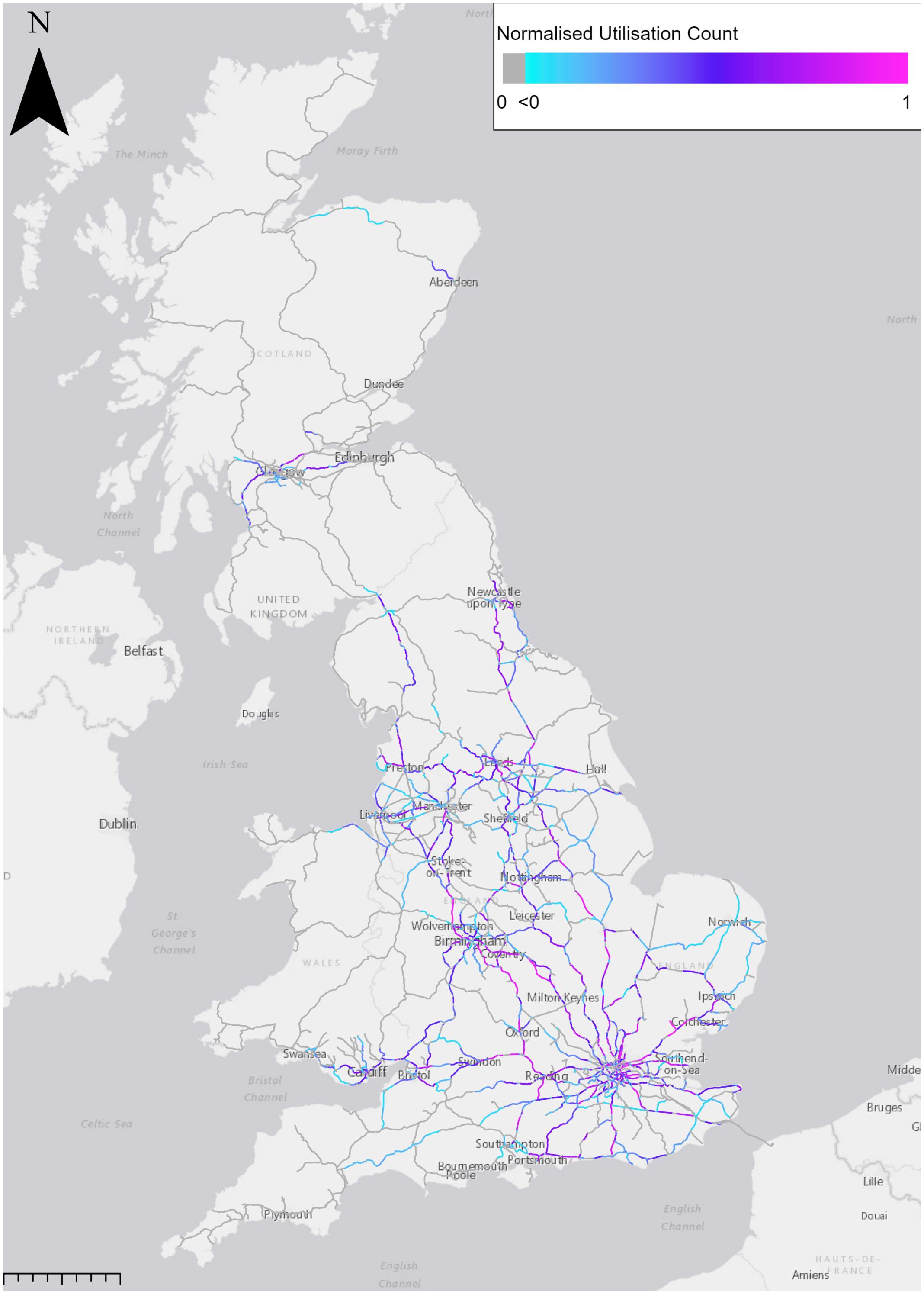


Figure C.20: Normalised Actual Utilisation for 17/07/2023

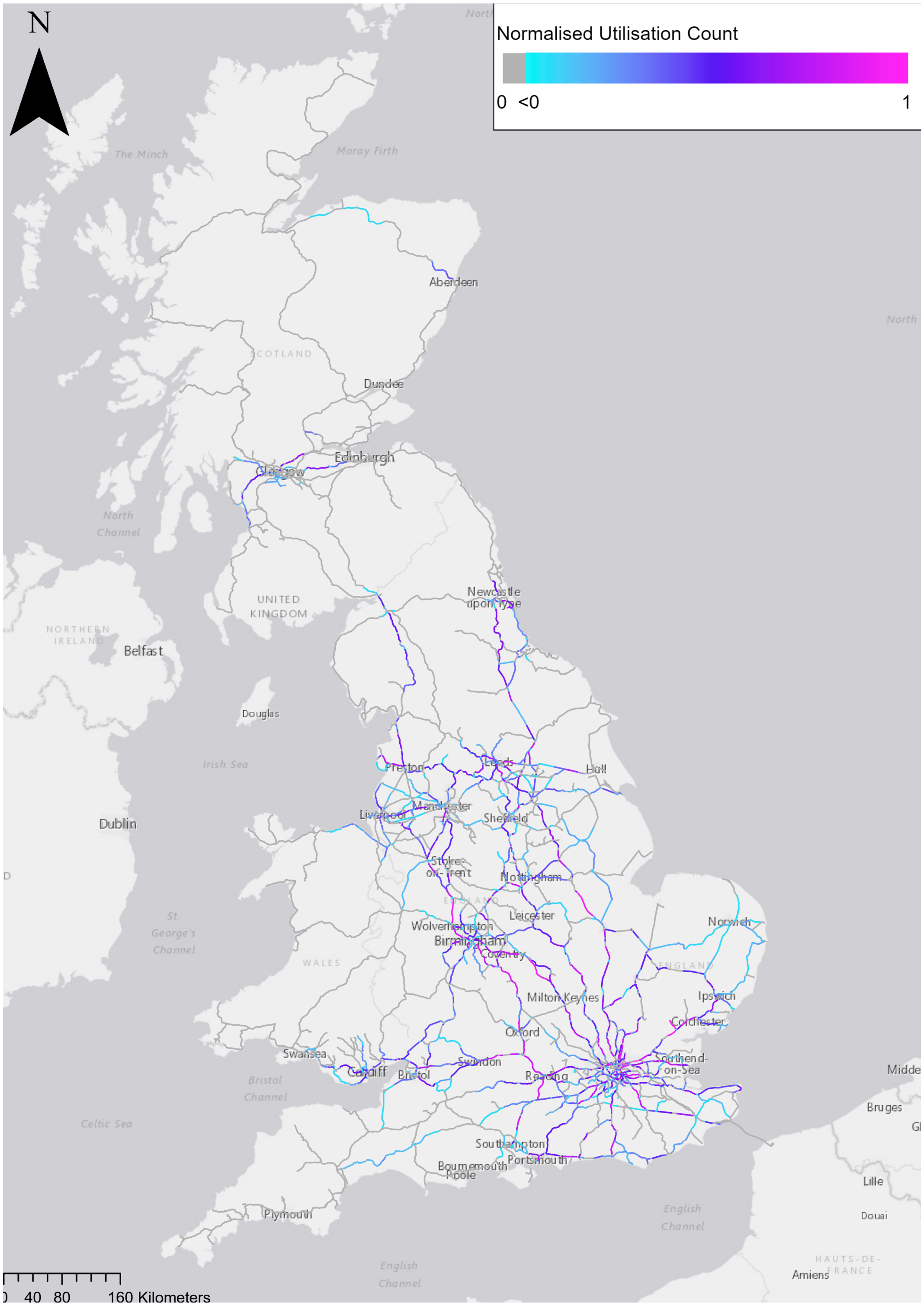


Figure C.21: Normalised Actual Utilisation for 18/07/2023

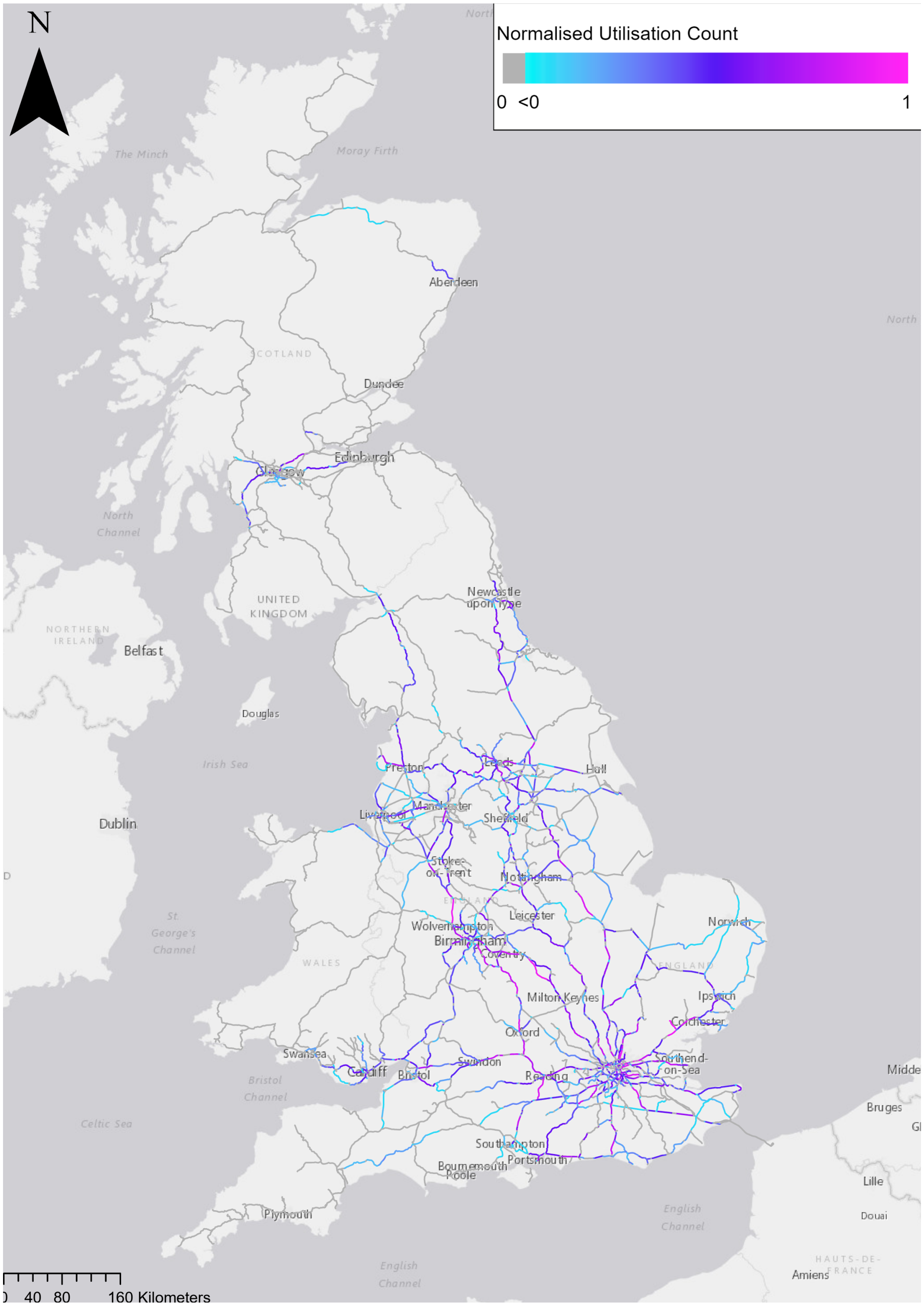


Figure C.22: Normalised Actual Utilisation for 19/07/2023

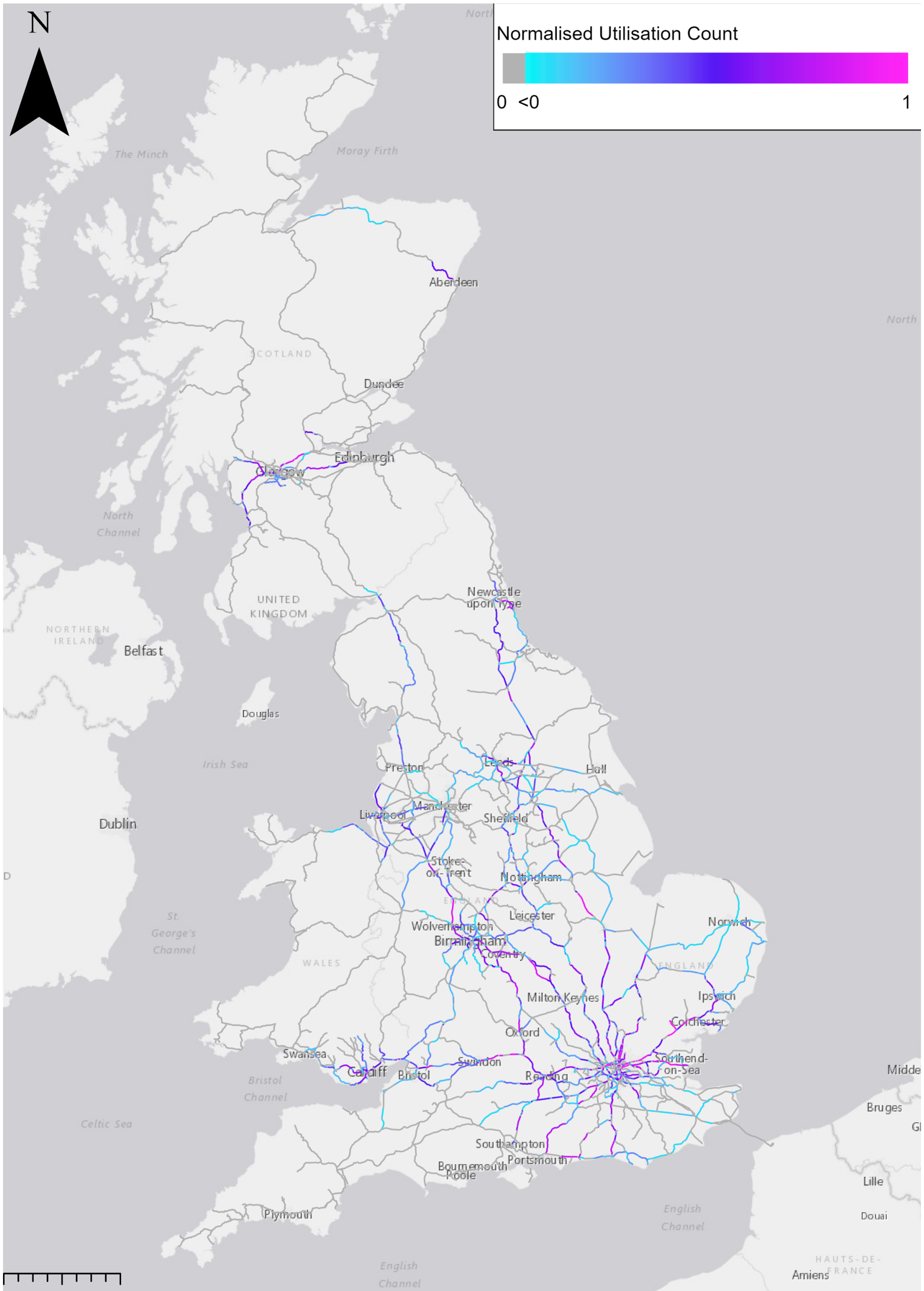


Figure C.23: Normalised Actual Utilisation for 20/07/2023

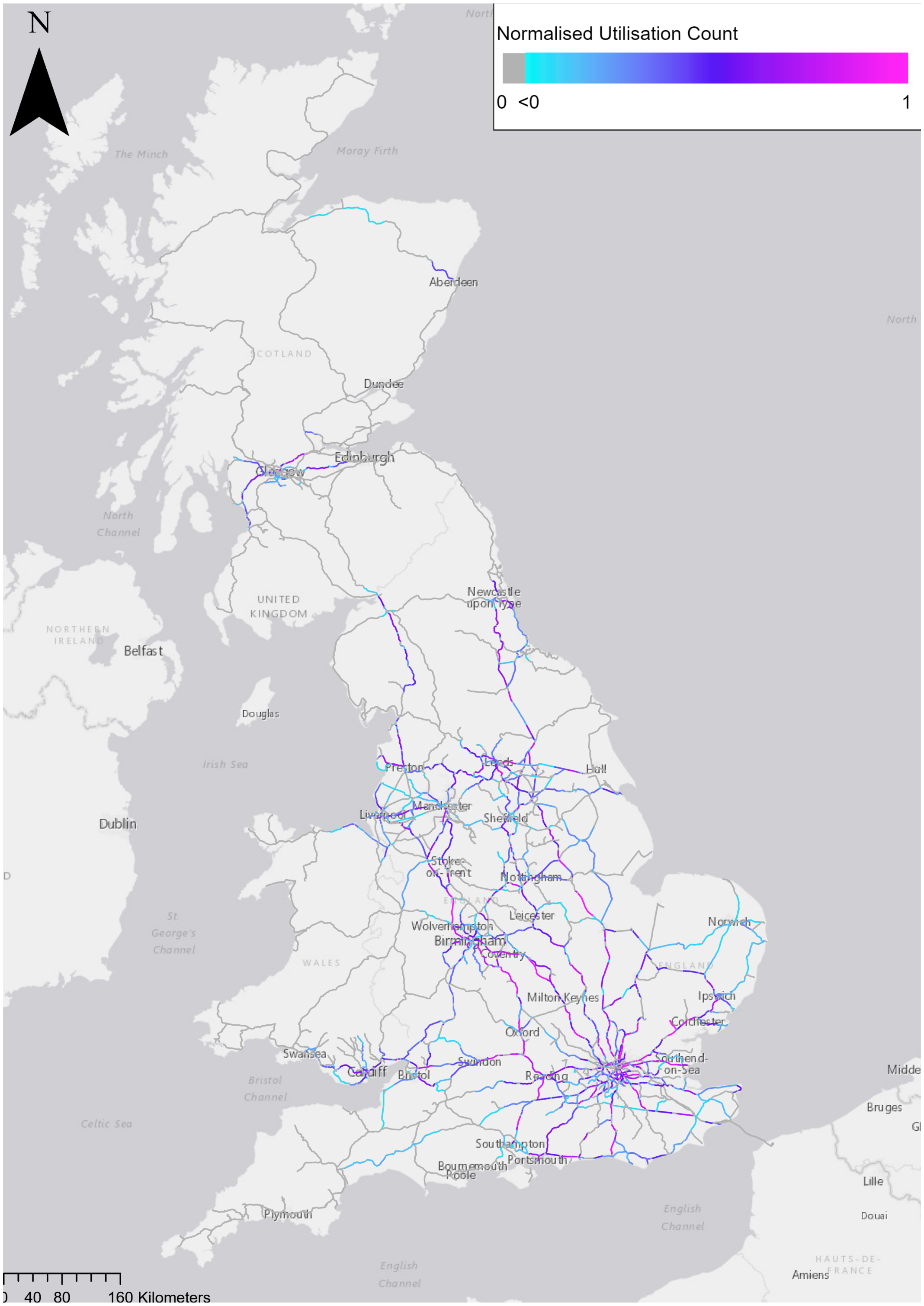


Figure C.24: Normalised Actual Utilisation for 21/07/2023

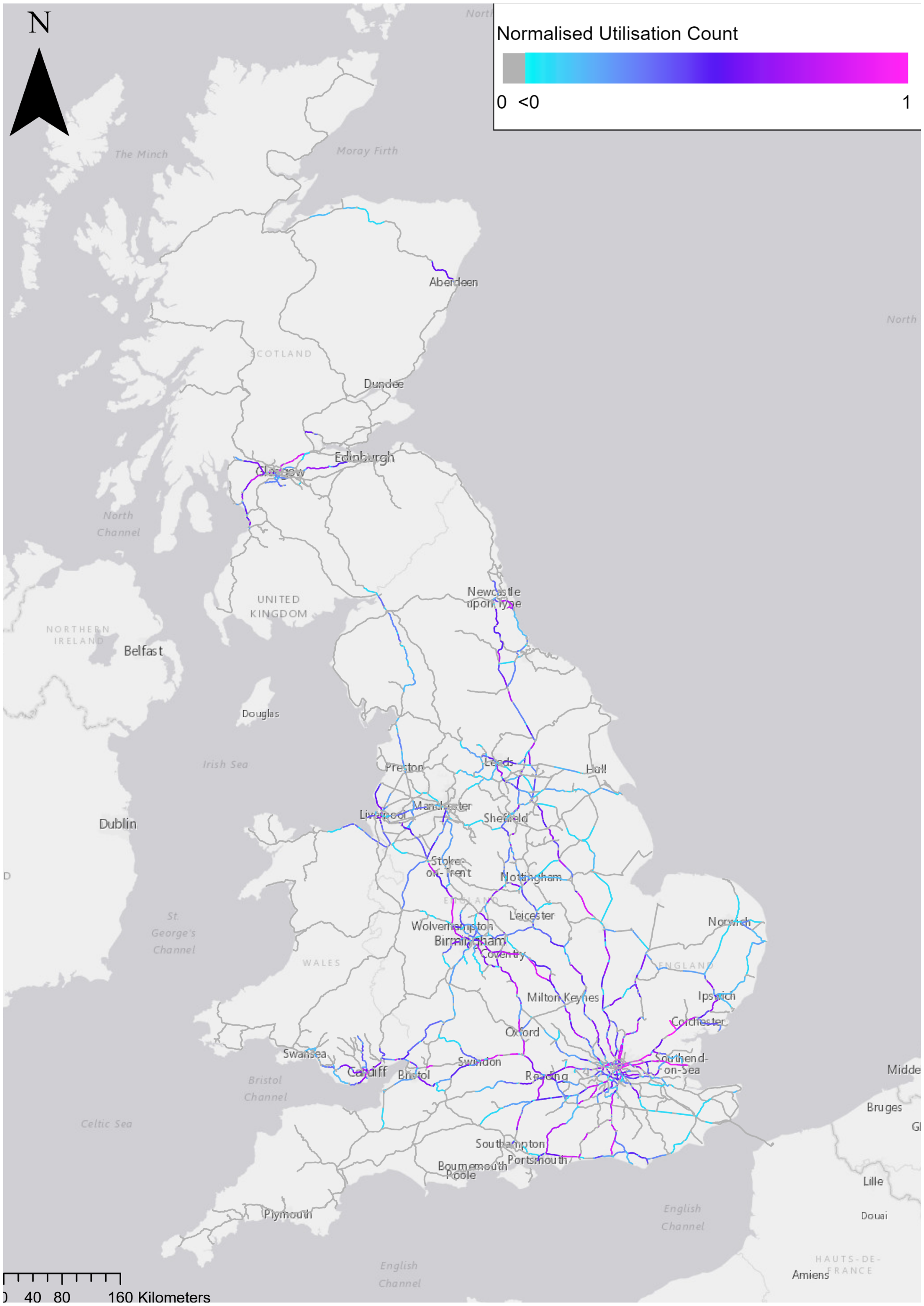


Figure C.25: Normalised Actual Utilisation for 22/07/2023

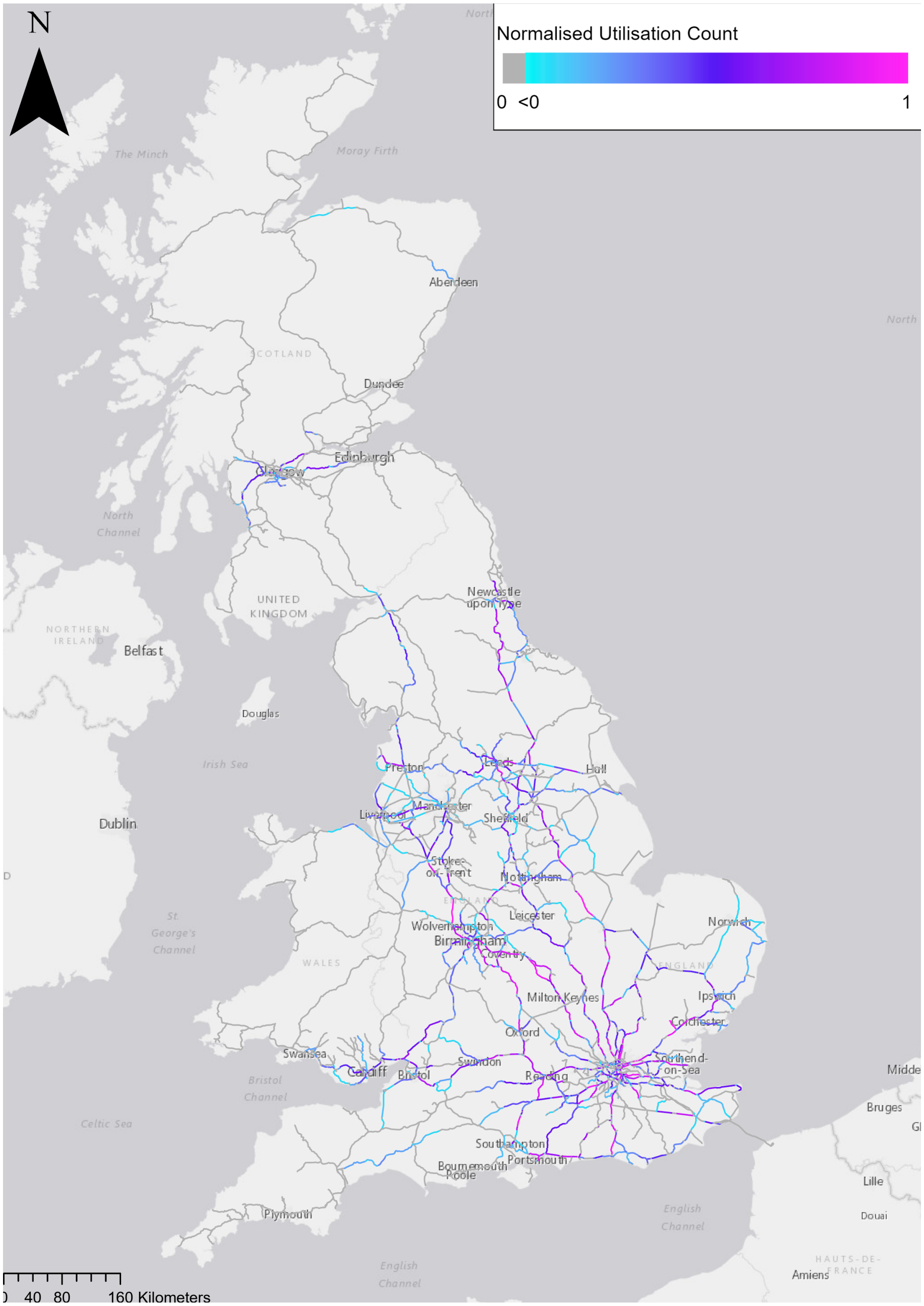


Figure C.26: Normalised Actual Utilisation for 23/07/2023

Figure C.27: Difference between Planned and Actual Utilisation for 11/07/2023

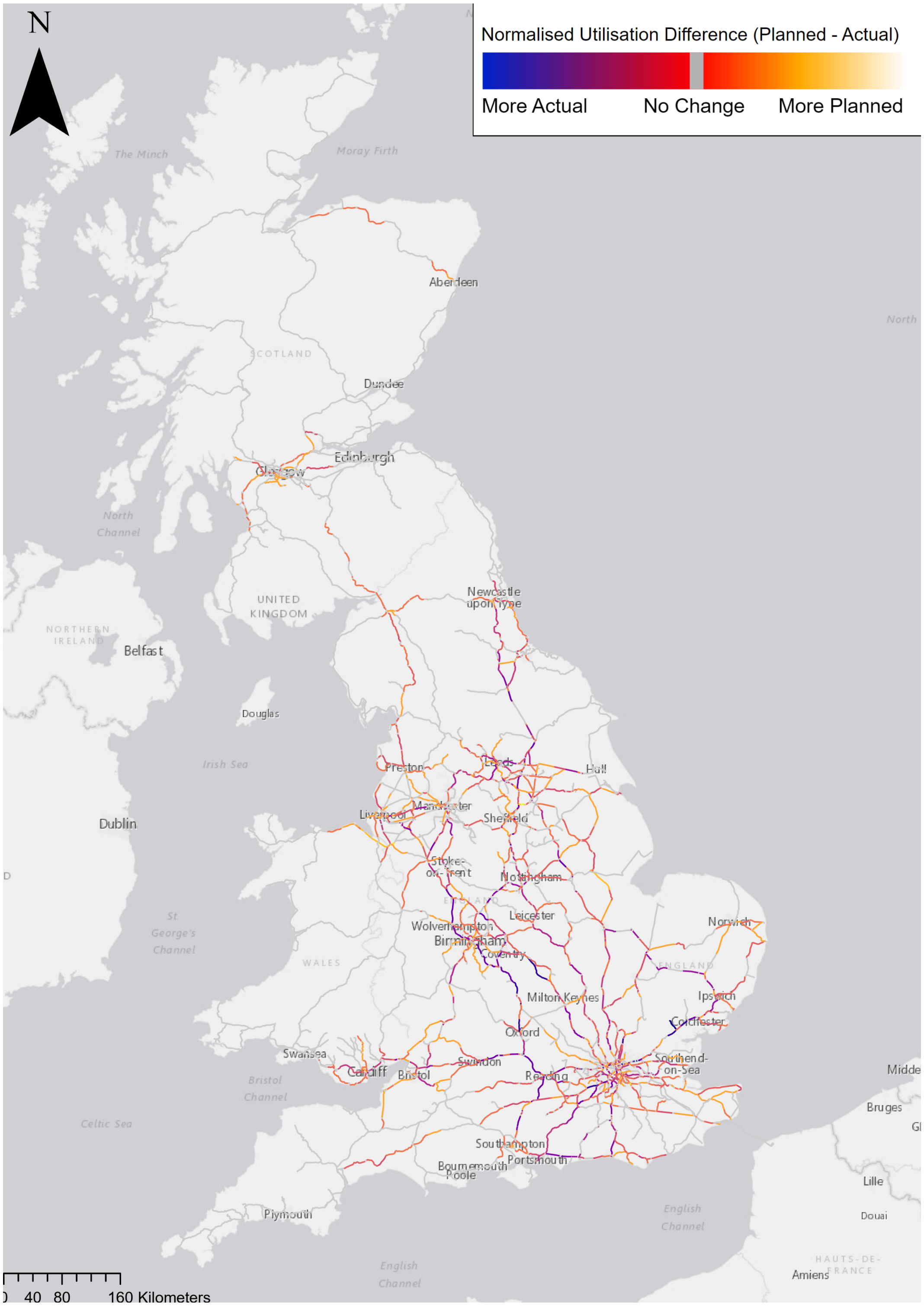


Figure C.28: Difference between Planned and Actual Utilisation for 12/07/2023

Figure C.29: Difference between Planned and Actual Utilisation for 13/07/2023

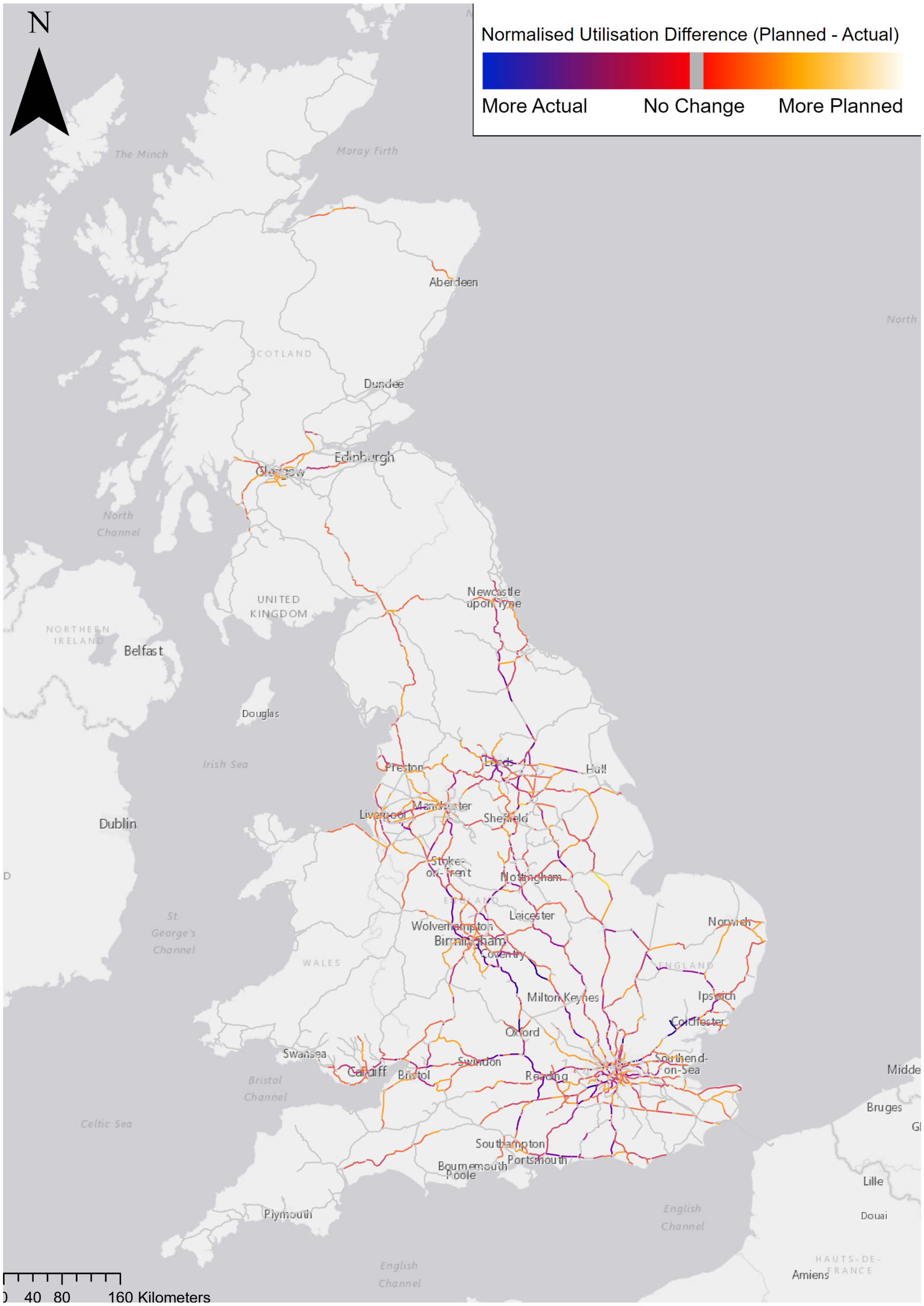


Figure C.30: Difference between Planned and Actual Utilisation for 14/07/2023

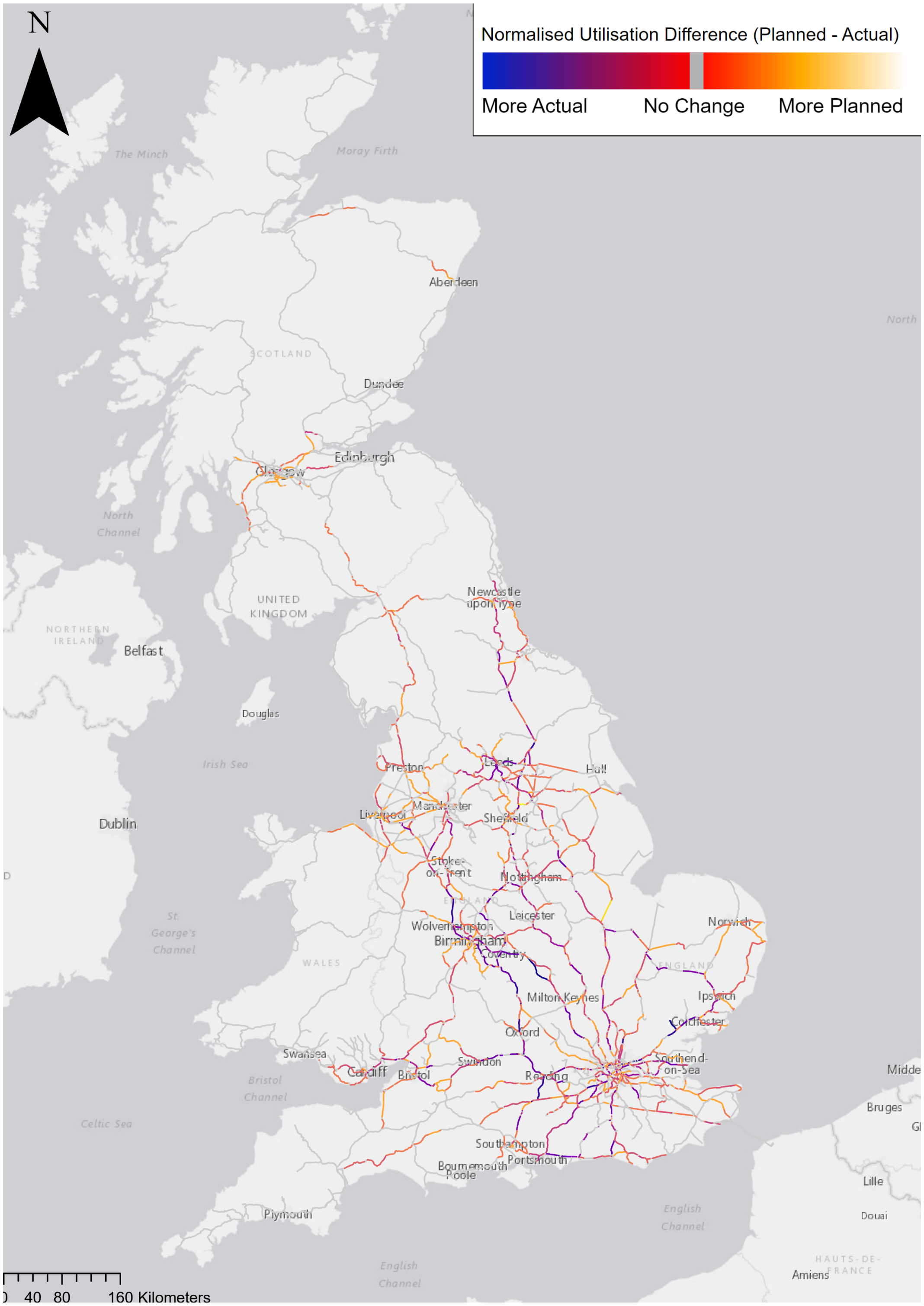


Figure C.31: Difference between Planned and Actual Utilisation for 15/07/2023

Figure C.32: Difference between Planned and Actual Utilisation for 16/07/2023

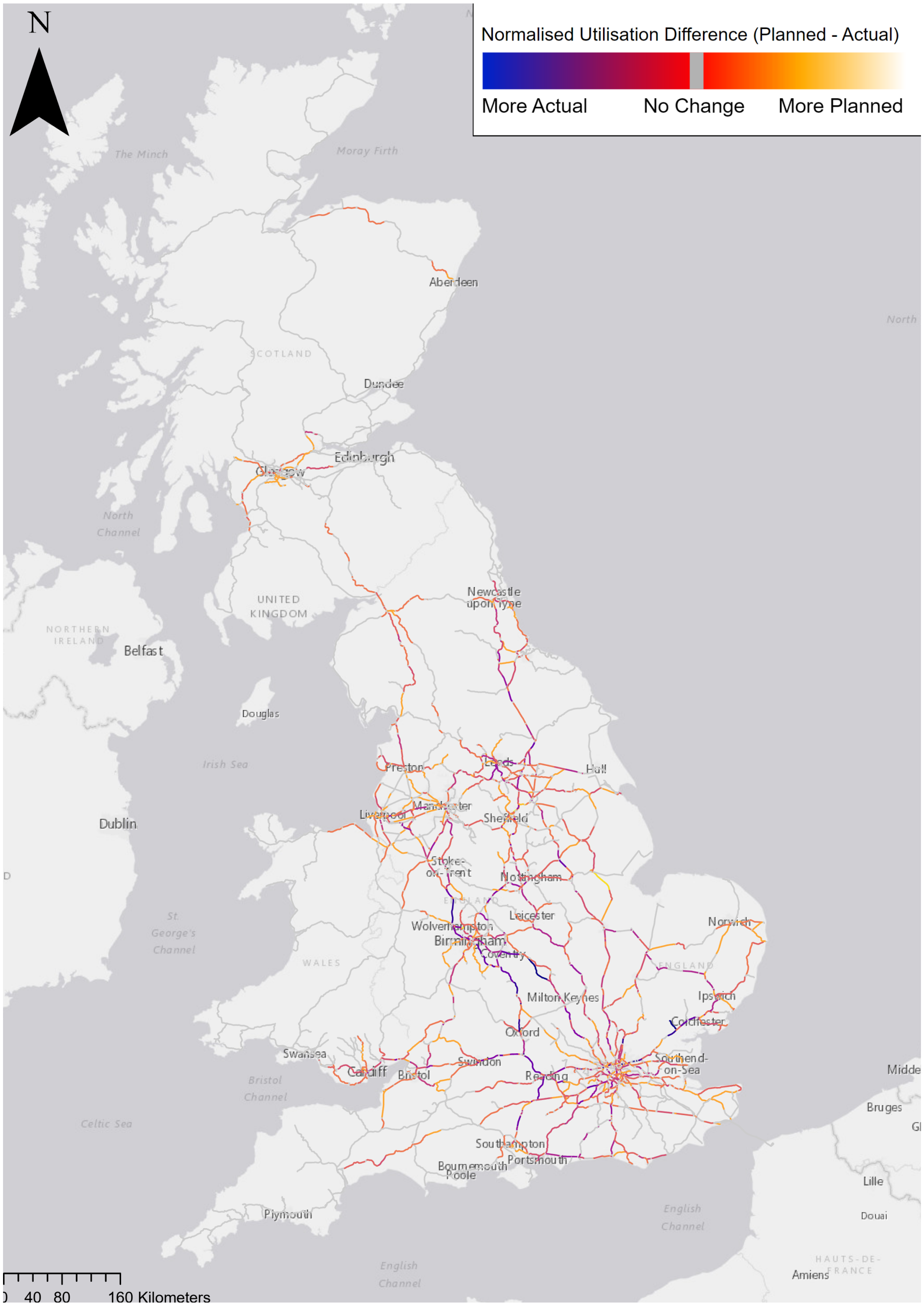


Figure C.33: Difference between Planned and Actual Utilisation for 17/07/2023

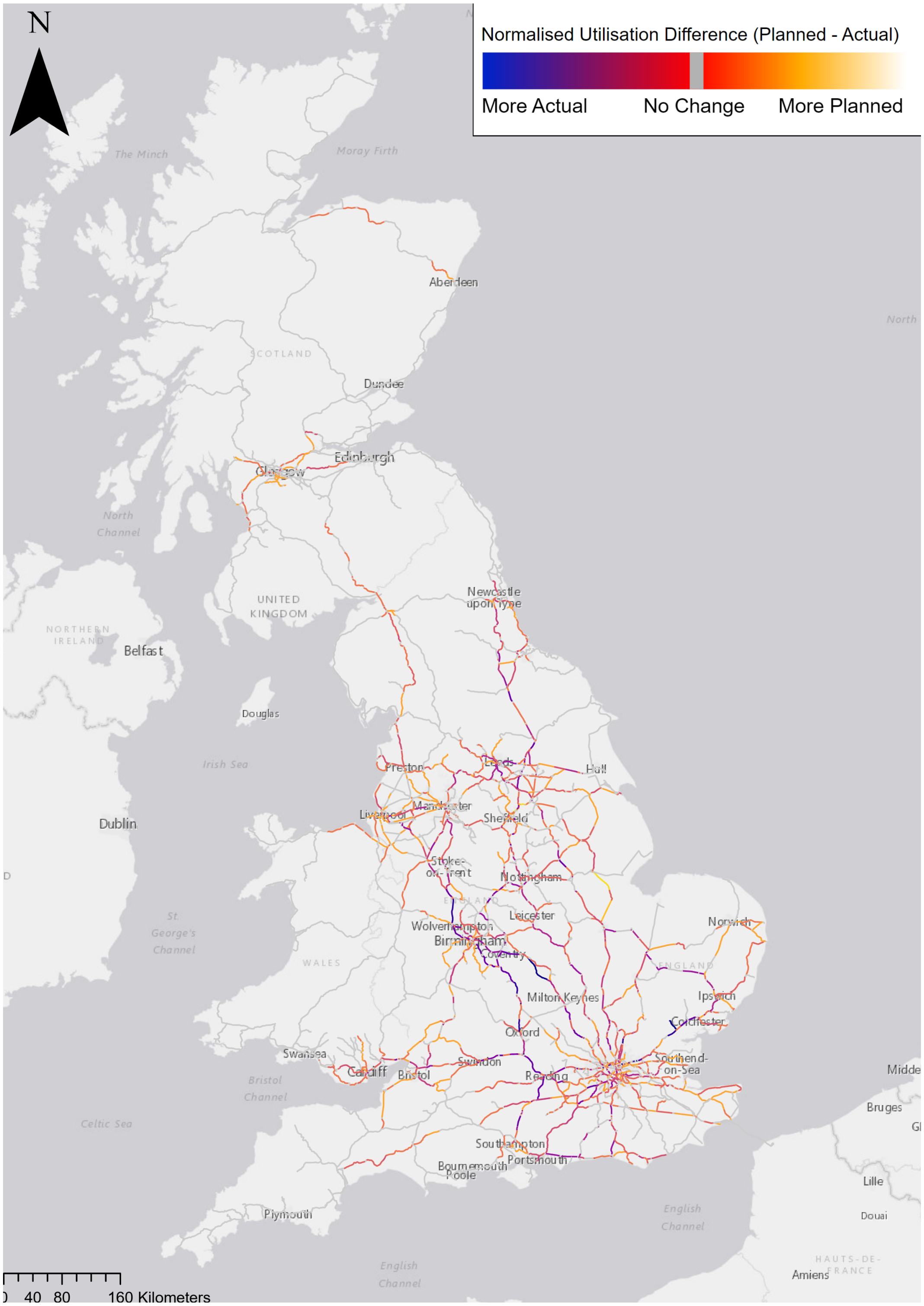


Figure C.34: Difference between Planned and Actual Utilisation for 18/07/2023

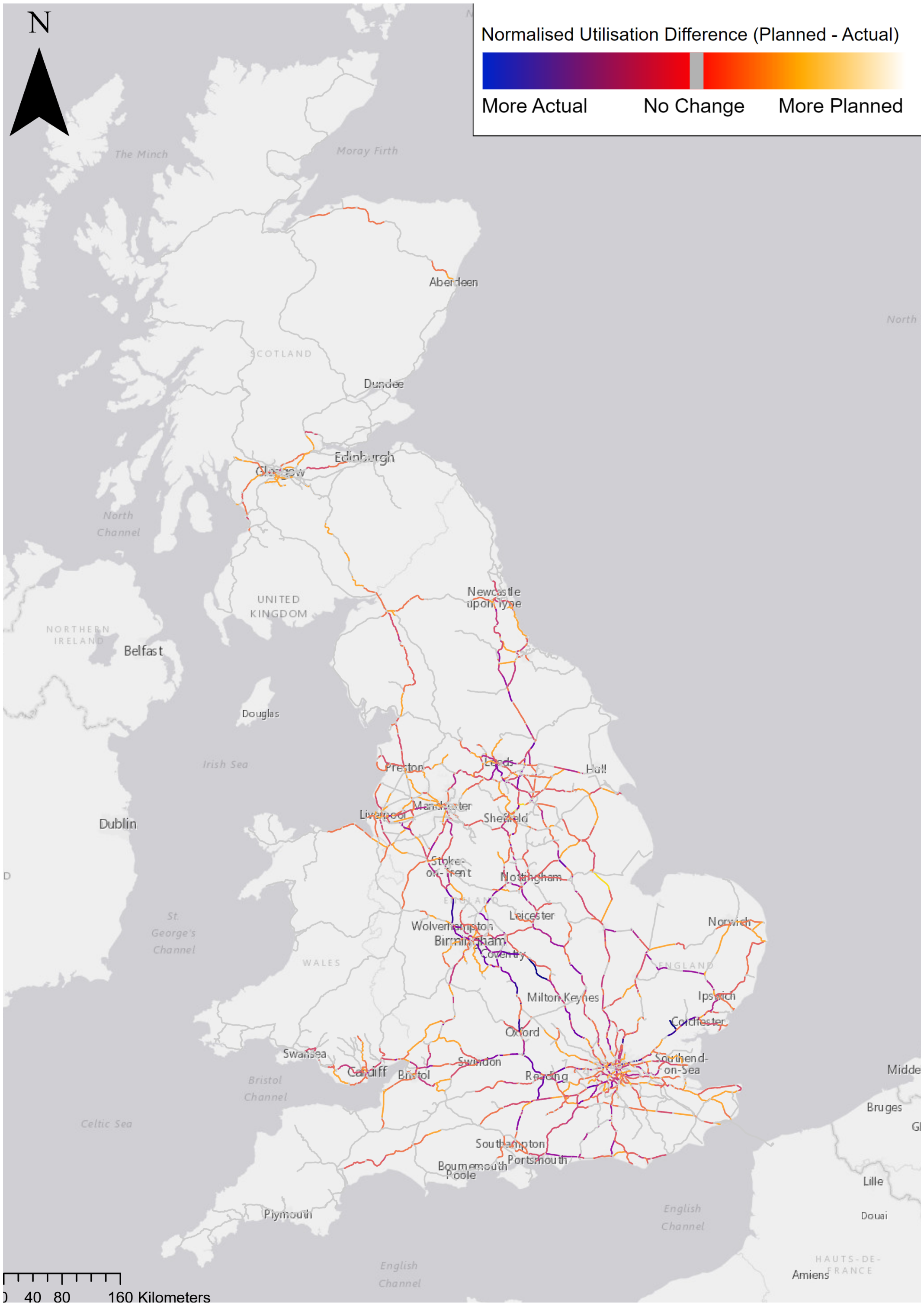


Figure C.35: Difference between Planned and Actual Utilisation for 19/07/2023

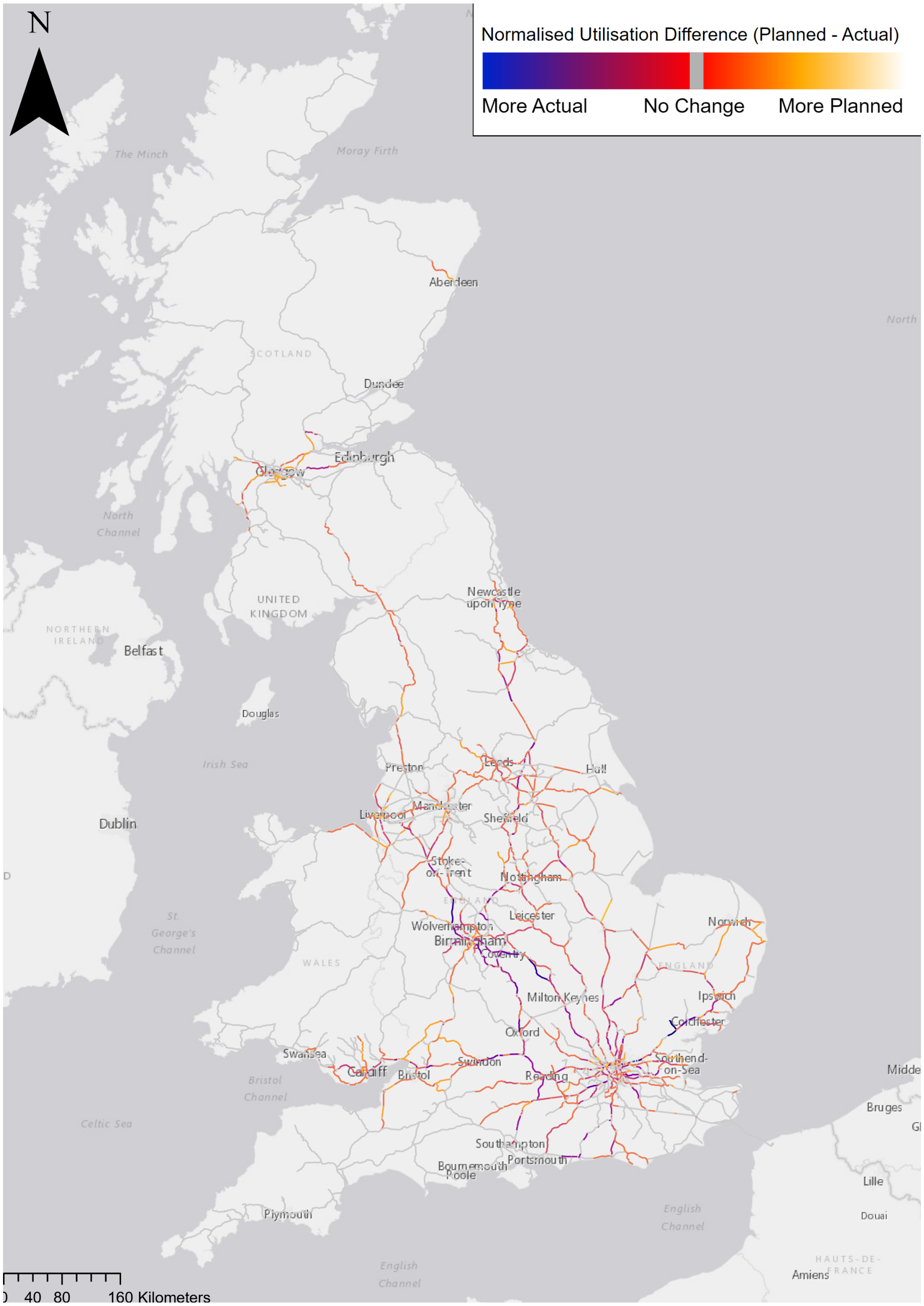


Figure C.36: Difference between Planned and Actual Utilisation for 20/07/2023

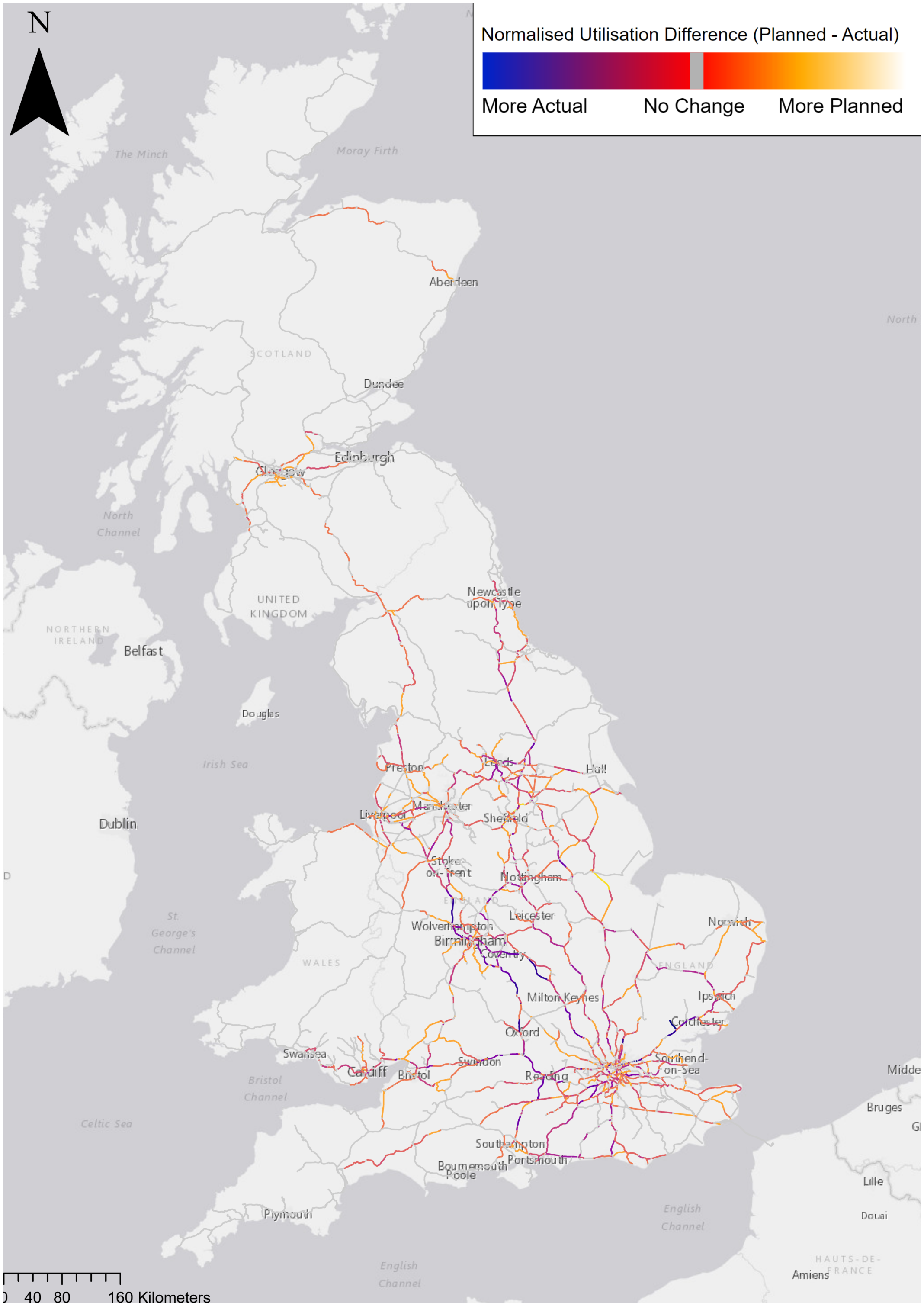


Figure C.37: Difference between Planned and Actual Utilisation for 21/07/2023

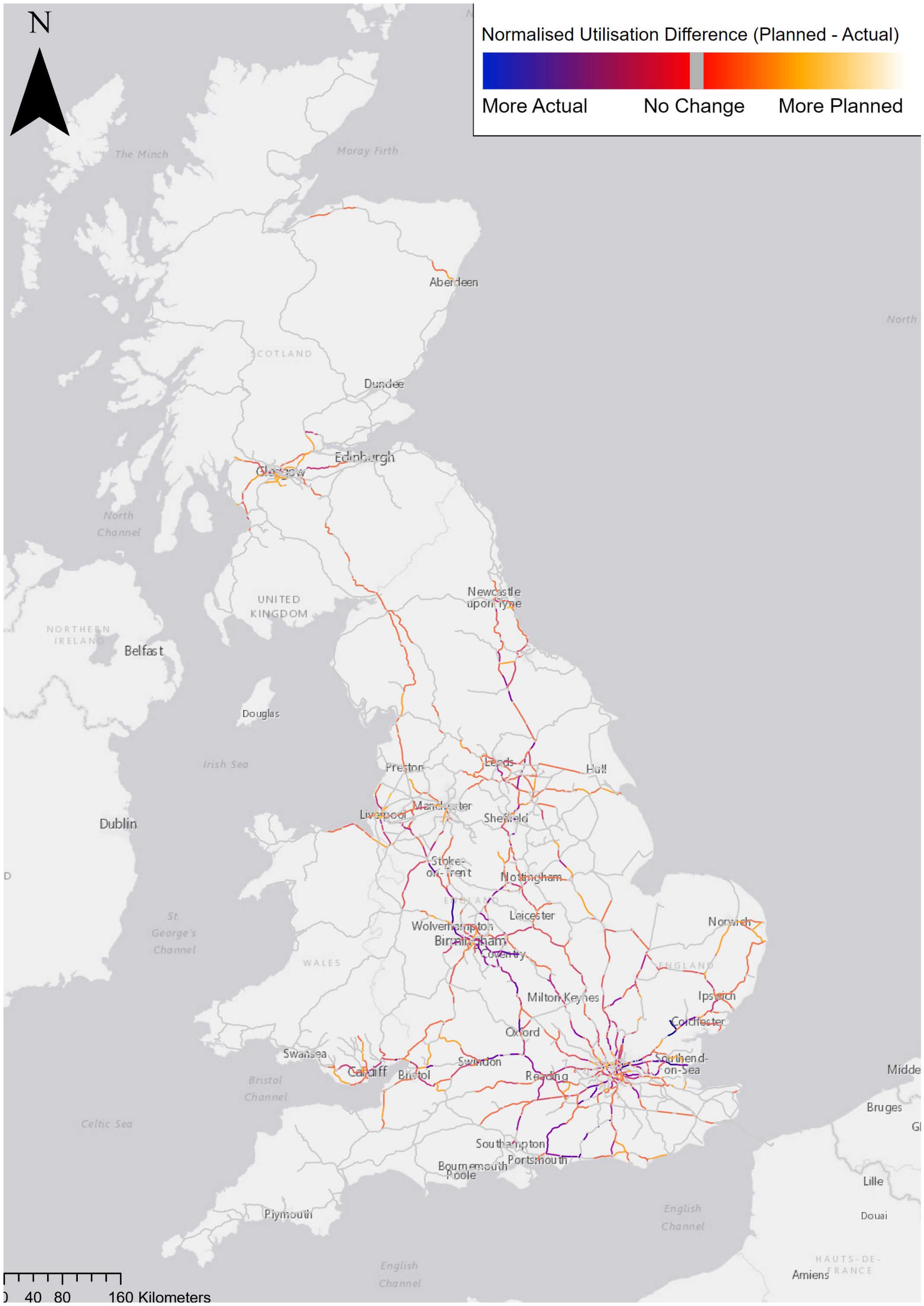


Figure C.38: Difference between Planned and Actual Utilisation for 22/07/2023

Figure C.39: Difference between Planned and Actual Utilisation for 23/07/2023