



Spatiotemporal Analysis and Prediction of Pipe Failures in a Water Distribution Network Using Cluster Analysis and near and Spatial Join Geoprocessing Tools [†]

Zoi Papavasileiou ^{1,*} and Vasilis Kanakoudis ²

¹ Civil Engineering Department, University of Thessaly, 38334 Volos, Greece

² Civil Engineering Department, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; kanakoudis@civil.auth.gr

* Correspondence: zpapav@gmail.com

[†] Presented at the 6th International Conference on Efficient Water Systems (EWaS6), Thessaloniki, Greece, 11–14 May 2026.

Abstract

Water loss and significant problems in the operation of water distribution networks caused by pipe failures are a global problem that needs immediate attention. This study is based on the experience-based assumption that the probability of water main breaks occurring is highest within a short time and a short distance from a previous (considered initial or base) break. The dataset used includes the historical pipe breaks recorded from 2007 to 2020 in the city of Larisa, Greece. A Geographic Information System (GIS) application is used for better data visualization, but also for effective operation and management of the developed water network database. Cluster analysis and Near and Spatial Join geoprocessing tools are the main tools used to detect and analyze trends in data related to space and time. In addition, the study attempts to identify relations between pipe attributes (material, age), environmental stressors (traffic load, soil type), and spatiotemporal clustering patterns. Finally, a machine learning-based water pipe failure Prediction Model is developed to serve as the computational engine of a Decision Support System (DSS) designed to optimize pipe replacement prioritization.

Keywords: pipe; breaks; spatiotemporal; prediction; GIS

1. Introduction

Water Distribution Networks (WDNs) are critical urban infrastructures that ensure reliable delivery of potable water to their customers and are essential for public health and economic security. Structural deterioration of these assets, particularly pipe failures, bursts, and leaks, poses significant challenges, leading to service interruptions, water loss, costly repairs, and substantial financial burdens for water utilities. In Greece, as in many parts of the world, Municipal Water and Sewerage Companies (DEYAs) have traditionally relied on reactive maintenance strategies, addressing pipe failures only after they occur. High indirect costs and social impact associated with emergency repairs prove this “run-to-failure” approach unsustainable.

Addressing these challenges requires a transition toward proactive Asset Management (AM). A critical aspect of this transition is the capacity to predict future failure events. This study is based on the experience-based conclusion that pipe failures are not random events but exhibit spatiotemporal correlation; specifically, the probability of a water main break is



Academic Editors: Evangelos Keramaris, Francesco De Paola, Habib Muhammetoglu and Theodora Papamitsou

Published: 23 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

highest within a short temporal and spatial interval following a previous (original/base) failure event [1]. This actually works as a space–time failure magnet.

The current study introduces a comprehensive methodology applied to the WDN of Larisa, Greece. The objectives of utilizing Geographic Information Systems (GIS) and Machine Learning (ML) tools, are to (a) visualize and manage network data efficiently; (b) employ advanced spatial statistical tools such as Cluster Analysis and Near and Spatial Join geoprocessing tools to identify trends and enhance the theory of time–space correlation of pipe breaks; and (c) develop a predictive model that functions as the core of a Decision Support System (DSS) for the optimization of the pipe replacement prioritization.

2. Materials and Methods

The proposed methodology follows a three-step workflow: (a) Data Pre-processing; (b) Spatiotemporal Analysis; and (c) Vulnerability Modeling.

2.1. Case Study and Data Acquisition

The study area is the city of Larisa, served by the Municipal Water and Sewerage Company of Larisa (DEYAL, Larisa, Greece). A robust geodatabase was developed in the ArcGIS Pro (Version 3.6.2) environment to consolidate heterogeneous data sources. The primary dataset consists of:

- **Historical Failure Records:** A comprehensive log of pipe breaks recorded over a 14-year period, from 2007 to 2020 (Figure 1). The 2300 pipe breaks analyzed in this study refer strictly to failures, such as structural pipe body breaks and joint leaks, in the main distribution network pipelines; all recorded maintenance events concerning consumer service connections, hydrants, and secondary fittings were manually isolated and excluded from the dataset.

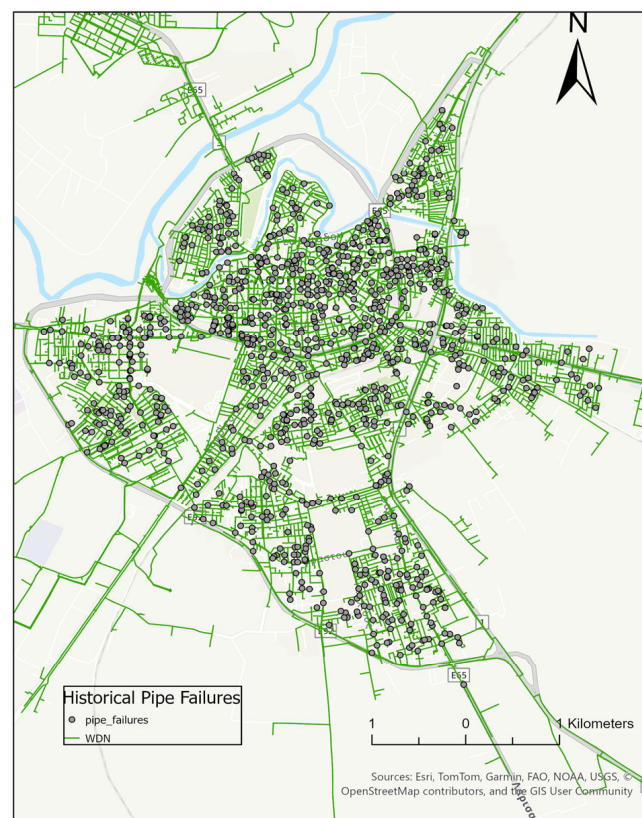


Figure 1. Map of Larisa Water Network & Failure Points 2007–2020.

- **Network Infrastructure:** Spatial data representing the water mains, including attributes such as material, age, and diameter.
- **Environmental Stressors:** Spatial layers representing traffic load (derived from the road network hierarchy) and soil type, which are considered critical external factors influencing pipe integrity.

2.2. Spatiotemporal Analysis-Theoretical Framework

To validate the assumption that breaks are clustered in space and time (space–time failure magnet phenomenon), specific GIS tools were employed:

- **Cluster Analysis:** Used to identify areas with high failure density, providing an initial overview of the network’s vulnerability [2]. Initially, a spatiotemporal density-based cluster analysis, utilizing the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, was applied to the historical failure dataset. In this research, this clustering algorithm transcends static spatial boundaries to identify baseline vulnerability zones by simultaneously evaluating spatial and temporal proximity. By calculating the distance of each pipe break to its minimum required neighboring failures within strictly defined spatiotemporal windows—specifically 10 m within 10 days (capturing the immediate space–time failure magnet effect) and 50 m within 30 days (capturing the broader neighborhood scenario)—the algorithm effectively delineates dense spatiotemporal groupings (clusters) of repair events across the entire 2007–2020 period, filtering out isolated, random breaks as spatial noise. This method highlights specific network areas where intrinsic factors and the subsequent “space–time magnet effect” systematically compromise pipe integrity, leading to failures that cluster tightly in both space and time.
- **Near and Spatial Join geoprocessing tools:** They were used to calculate the exact physical distance of each pipe segment to the nearest historical break. Concurrently, to fully capture the temporal dimension of the Goulter & Kazemi assumption [1], the time elapsed between consecutive failure events was computed. This dual approach allowed the model to identify a “space–time magnet” effect.
- **Python script:** To optimize the workflow across the extensive 14-year dataset, a custom Python script using the ArcPy library in ArcGIS Pro (Version 3.6.2) was also developed. This programmatic approach automated the distance calculations and concurrently computed the exact time elapsed between consecutive failure events.

By combining manual geoprocessing validation with advanced automated Python evaluation, this methodology allowed the model to efficiently identify a “space–time failure magnet” effect in its entirety in order to confirm that subsequent breaks are highly probable to occur not only within a tight spatial radius but also within a critically short time window following the initial repair.

2.3. Machine Learning Prediction and DSS

The final stage involves the development of a failure prediction model using Machine Learning algorithms (e.g., Random Forest) [3]. The model is trained to recognize complex, non-linear relationships between the input variables (pipe attributes, environmental stressors, and spatiotemporal cluster indices) and the occurrence of pipe breaks [4]. The output of this model—the Probability of Failure (PoF)—serves as the computational engine for a Decision Support System (DSS). This DSS allows the utility to move beyond simple age-based replacement and adopt a risk-based prioritization strategy [5].

3. Results and Discussion

The application of the proposed methodology to the Larisa WDN yielded significant insights regarding both the historical pipe failure patterns and the predictive capabilities of the developed model.

3.1. Spatiotemporal Analysis

The statistical analysis, using the tools provided by ArcGIS Pro (Version 3.6.2), of the 2300 failure records from 2007 to 2020, revealed distinct patterns in the occurrence of pipe breaks.

To empirically validate the Goulter and Kazemi assumption, which posits that pipe failures are spatiotemporally clustered and heavily influenced by nearby prior breaks [1], a two-tier GIS methodology was employed.

3.1.1. Cluster Analysis

Macroscopically, the network's general spatiotemporal clustering was verified using Space–Time Pattern Mining tools. To identify localized spatiotemporal patterns and understand how consecutive pipe failures group together within the Water Distribution Network (WDN) of Larisa, the Spatiotemporal Density-Based Spatial Clustering of Applications with Noise (ST-DBSCAN) algorithm was applied (Figure 2). To evaluate the sensitivity of the clustering formation and capture different operational scales of network deterioration, the algorithm was executed under three distinct parameter scenarios. The temporal (Δt) and spatial (Δs) search distances were gradually increased to observe the aggregation of failure events:

- **Scenario 1** (Strict Thresholds— Δs : 1 m, Δt : 1 day): Under the most restrictive parameters, assuming that related failures occur almost simultaneously and at the exact same location, the algorithm identified a total of 251 distinct clusters. This high number of micro-clusters primarily represents immediate, localized re-breaks or successive repair interventions on the exact same pipe segment within a 24 h window.
- **Scenario 2:** (Moderate Thresholds— Δs : 10 m, Δt : 10 days): Expanding the spatial search radius to 10 m and the temporal window to 10 days resulted in the consolidation of events into 92 clusters. This scenario captures short-term neighborhood vulnerabilities: an initial pipe burst may cause hydraulic instability or pressure fluctuations, leading to subsequent failures in adjacent areas within 10 days.
- **Scenario 3:** (Broader Thresholds— Δs : 50 m, Δt : 30 days): By setting the spatial tolerance to 50 m and the temporal threshold to 30 days, the algorithm merged the failure events into 73 major macro-clusters. This broader grouping effectively highlights broader structural weaknesses in specific district metering areas (DMAs) or street blocks, reflecting extended phases of network degradation over a monthly period.

The progressive reduction in the total number of clusters (from 251 to 73) as the spatiotemporal thresholds expanded highlights the hierarchical nature of failure propagation in the WDN of Larisa. While strict parameters isolate specific, acute repair events, broader parameters successfully reveal wider “vulnerability zones” that require systematic pipe replacement rather than spot repairs.

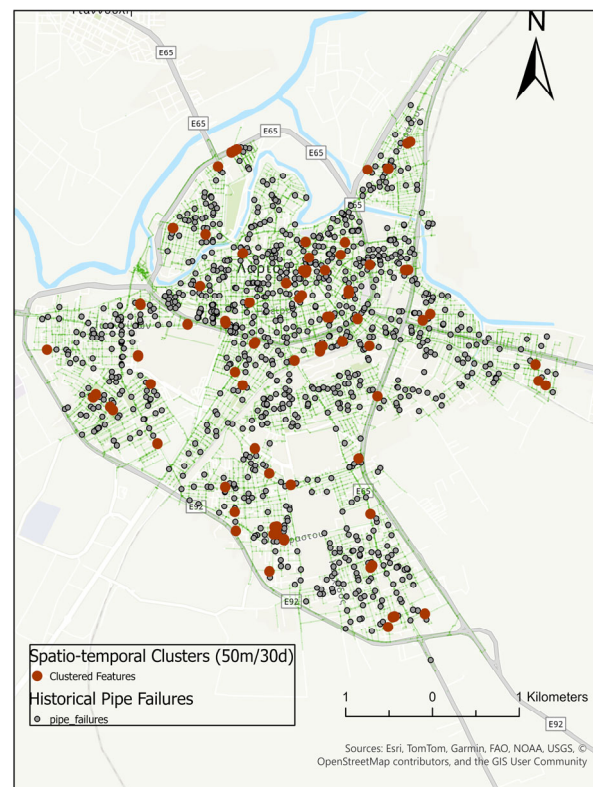


Figure 2. Spatiotemporal clustering of pipe failures (DBSCAN algorithm).

3.1.2. Validation of the Goulter & Kazemi Assumption (Time–Space Clustering)

Microscopically, the direct causal relationship of the “space–time magnet effect” was quantified. The Near (and Spatial Join) geoprocessing tools, combined with a custom Python script using the ArcPy library in ArcGIS Pro, were used to calculate the exact physical distance of each pipe segment to the nearest historical break (NEAR_DIST). The investigation into the Goulter and Kazemi hypothesis confirmed the existence of spatiotemporal clustering [1]. Specifically, it was observed that consecutive pipe failures exhibit a strong tendency to occur within tight spatial and temporal windows following an initial break event.

The quantitative spatial analysis revealed that 31.07% of the recorded failures (715 events) occurred within a highly restricted radius of 1 m and within a 1-day temporal window of a previous break. This essentially represents immediate re-breaks or successive, localized repair interventions at the exact same location. Expanding the spatial radius slightly to 20 m within the same 1-day timeframe marginally increased this proportion to 31.25% (719 events).

As the spatiotemporal thresholds broadened, the magnet effect became progressively more pronounced. Within a 10 m radius and a 10-day window, 46% of total failures (1.059 events) had a spatiotemporal “neighbor”, validating the hypothesis of short-term neighborhood vulnerability induced by hydraulic pressure fluctuations following an initial burst. Ultimately, extending the analysis to a 50 m spatial tolerance and a 30-day temporal window demonstrated that nearly half of the network’s total breaks (49%, or 1.128 events) are spatiotemporally correlated (Table 1).

These findings indeed validate the space–time clustering hypothesis for the Larisa WDN. They demonstrate that a primary pipe failure acts as a space–time magnet, drastically increasing the probability of subsequent structural failures in its immediate vicinity and in the near future. Recognizing this high correlation is critical to transitioning from reactive spot repairs to proactive, block-level pipe replacements.

Table 1. Spatiotemporal clustering of consecutive pipe failures based on varying spatial (Δs) and temporal (Δt) thresholds.

Scenario	Spatial Window (Δs)	Temporal Window (Δt)	Percentage of Total Breaks (%)
Immediate/Rebreaks	≤ 1 m	≤ 1 day	31%
Localized Short-term	≤ 10 m	≤ 10 days	46%
Broad Systemic	≤ 50 m	≤ 30 days	49%

Ultimately, these quantitative findings provide robust empirical evidence validating the foundational hypothesis proposed by Goulter and Kazemi regarding the spatiotemporal clustering of water main breaks [1]. The data from the Larisa WDN mathematically confirms their assertion that a primary pipe failure is not an isolated incident, but rather acts as a localized catalyst—a “space–time magnet”—that drastically increases the probability of subsequent structural failures in its immediate physical vicinity and near future. Recognizing that breaks are spatiotemporally correlated dictates a critical shift in asset management: moving away from reactive, localized spot repairs and toward proactive, block-level pipe-replacement strategies.

3.2. Machine Learning Model Performance

A multivariate spatial analysis was conducted in order to comprehensively evaluate the mechanisms driving water pipe failures. The examined explanatory variables encompassed historical spatiotemporal metrics (Proximity to prior failures), physical pipe attributes (Length, Diameter, Age, Material), environmental and external stressors (Traffic Load, Climatic Conditions, Land Use, Proximity to other utility trenches) [6]. The requisite spatial relationships were initially constructed using the Near and Spatial Join geoprocessing tools in ArcGIS Pro, which calculated the exact physical distances and topological intersections for the failed segments.

To quantify the relative impact of each parameter and generate the Feature Importance ranking, the Forest-based Classification and Regression (an advanced machine learning algorithm within the ArcGIS Pro Spatial Statistics toolbox) was employed [7]. To ensure unbiased learning, the inherent data imbalance between historical failure records and intact pipes was efficiently addressed using the algorithm’s native sampling and weighting mechanisms, without the need for synthetic data generation. Consequently, the model was trained on a statistically representative sample of the actual Larisa network, accurately incorporating derived spatiotemporal thresholds such as failure clustering at near distance and high failure density in short segments. Finally, the model’s predictive accuracy and robustness were successfully validated utilizing the internal Out-of-Bag (OOB) error estimation.

The Feature Importance ranking (Table 2) identified spatial proximity to a previous break as the dominant predictive factor (29.16%), mathematically validating the Goulter and Kazemi “domino effect” hypothesis [1]. A secondary group of variables followed with balanced contributions: network age (14.27%), pipe segment length (14.22%), traffic load (13.94%), pipe diameter (13.63%), and climate (13.61%). Ultimately, these results indicate that while localized stresses from recent nearby repairs act as the primary failure catalysts, pipe degradation remains a multi-causal phenomenon significantly influenced by inherent network characteristics and external operational loads.

Based on these predictions, a Vulnerability Map was generated that classified each pipe segment into risk categories (Low, Medium, High). This map serves as the visual interface of the Decision Support System (DSS).

Table 2. Random Forest Feature Importance ranking for pipe failure prediction.

Rank	Explanatory Variable (Failure Factor)	Importance (%)
1	Proximity to Previous Failure	29.16%
2	Network Age	14.27%
3	Pipe Segment Length	14.22%
4	Traffic Load (Road networks)	13.94%
5	Pipe Diameter	13.63%
6	Climatic Conditions	13.61%
7	Proximity to Other Utility Networks	0.52%
8	Pipe Material	0.33%
9	Land Use (Residential zones)	0.31%

3.3. Decision Support System (DSS) and Risk Prioritization Map

Following the validation of the Machine Learning model's predictive performance, the raw failure probabilities were integrated into a geospatial Decision Support System (DSS) to translate analytical data into actionable asset management strategies. Figure 3 illustrates the ultimate output of this framework: the Predicted Failure Risk map for the Larisa Water Distribution Network.


Figure 3. Spatial distribution of predicted failure risk across the Larisa WDN based on the proposed DSS framework.

This spatial visualization represents the final risk assessment (RISK_FINAL), which is mathematically derived by combining the model-generated Probability of Failure (PoF) for each pipe segment with its physical and operational Consequence of Failure (CoF) ($\text{Risk} = \text{PoF} \times \text{CoF}$). In this study, CoF is quantified using the pipe diameter, which serves as a direct proxy for the hydraulic impact of a potential break (i.e., larger-diameter mains receive higher CoF scores due to greater water losses and wider service disruptions). To

facilitate decision-making for the municipal authority (DEYAL), the network segments were categorized into three distinct operational tiers:

- **Low Risk (Blue):** Representing the majority of the network, these segments exhibit structural stability and require no immediate interventions, allowing for standard routine monitoring.
- **Moderate Risk (Yellow):** These pipes show early signs of degradation or are approaching the end of their optimal lifecycle. They require heightened surveillance and should be scheduled for medium-term reassessment.
- **High Risk/Priority Replacement (Red):** These critical segments are flagged for imminent failure due to a combination of high vulnerability (e.g., aging materials, historical spatiotemporal proximity to previous breaks) and severe operational consequences.

As observed in the spatial distribution, the high-risk segments are not uniformly scattered but are distinctly localized across specific network branches. Ultimately, the generation of this risk map serves as a cornerstone for DEYAL's strategic planning. It provides empirical evidence to support the transition from a reactive, "run-to-failure" maintenance paradigm to a proactive, targeted pipe replacement strategy. By prioritizing the "Red" zones, the utility can optimize its capital expenditure (CAPEX), prevent catastrophic water losses, and significantly minimize socio-economic disruptions within the urban fabric.

4. Conclusions

This study presented a holistic framework for transitioning from reactive to proactive asset management in the water utility sector. By utilizing the water network of Larisa, Greece, as a case study, we demonstrated that pipe failures are not random events but exhibit strong spatiotemporal correlations. The integration of GIS-based spatiotemporal analysis with Machine Learning (Random Forest) allowed for the accurate identification of high-risk zones. The key conclusions are:

- **Historical** failure data is a valuable asset that, when cleaned and georeferenced, can reveal hidden deterioration patterns.
- The "clustering" phenomenon is present in the Larisa network, necessitating post-repair inspections of neighboring pipes. The spatiotemporal analysis provided robust empirical evidence confirming the Goulter and Kazemi hypothesis [1] that a primary pipe failure acts as a space–time magnet for subsequent breaks.
- The **developed** predictive model serves as a reliable computational engine for a Decision Support System, enabling DEYAL to prioritize pipe replacements based on actual risk rather than just age. This facilitates a vital strategic shift: moving away from the traditional, reactive "run-to-failure" approach (which involves costly, repeated spot-repairs) toward a proactive, block-level pipe replacement strategy.

The main conclusion is that pipe failures in Larisa exhibit strong spatiotemporal clustering. While proximity to a previous break acts as the primary predictor of future failures, the model reveals that pipe degradation is a multi-causal phenomenon. This suggests a necessary shift toward risk-based, dynamic maintenance strategies that prioritize both spatial failure clustering and the physical deterioration of the network.

Future research could focus on expanding the DSS by integrating real-time hydraulic sensor data and dynamic pressure monitoring, further enhancing the real-time predictive capabilities of the framework.

Author Contributions: Conceptualization, V.K. and Z.P.; methodology, V.K. and Z.P.; software, Z.P.; validation, Z.P.; formal analysis, Z.P.; investigation, Z.P.; resources, Z.P.; data curation, Z.P.; writing—original draft preparation, Z.P.; writing—review and editing, V.K. and Z.P.; visualization, Z.P.; supervision, V.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions of the Water Utility (DEYAL).

Acknowledgments: To DEYAL (Larisa Water & Sewerage Company) for offering the historical pipe breaks data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Goulter, I.C.; Kazemi, A. Spatial and temporal groupings of water main pipe breakage in Winnipeg. *Can. J. Civ. Eng.* **1988**, *15*, 91–97. [CrossRef]
2. Chen, T.Y.-J.; Guikema, S.D. Prediction of water main failures with the spatial clustering of breaks. *Reliab. Eng. Syst. Saf.* **2020**, *203*, 107108. [CrossRef]
3. Xu, Y.; He, Z. Artificial Intelligence Prediction of Pipe Breaks in Water Distribution Systems Assisted by Spatial Clustering Analysis and Graph Neural Network. *ACS EST Eng.* **2026**, *6*, 647–657. [CrossRef]
4. Fan, X.; Wang, X.; Zhang, X.; Yu, X. Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors. *Reliab. Eng. Syst. Saf.* **2022**, *219*, 108185. [CrossRef]
5. Berrezel, Y.A.; Abdelbaki, C.; Rouissat, B.; Boumaaza, T.; Saber, M.; Goosen, M.; Kumar, N. Automated GIS-based decision tool for rehabilitation of a drinking water infrastructure in Tlemcen, Algeria. *Appl. Water Sci.* **2024**, *14*, 124. [CrossRef]
6. Robles-Velasco, A.; Cortés, P.; Muñuzuri, J.; De Baets, B. Prediction of pipe failures in water supply networks for longer time periods through multi-label classification. *Expert Syst. Appl.* **2023**, *213*, 119050. [CrossRef]
7. Esri. Forest-Based Classification and Regression (Spatial Statistics). ArcGIS Pro Documentation. Available online: <https://pro.arcgis.com/en/pro-app/3.4/tool-reference/spatial-statistics/forestbasedclassificationregression.htm> (accessed on 1 March 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.