FastAnalytics S.A.S

El futuro habla en datos

¡Lo anticipamos con decisiones!

Reconstructing spatio-temporal criminal networks for disruption in the Amazon

From partial communications to unobserved vertices and edges: mathematical construction, statistical evidence, and applied decision

Andrés Pérez-Coronado Gustavo Canavire-Bacarreza Santiago Ortíz Henry Laniado

FastAnalytics S.A.S · Scientific talk — mathematical-audience version

2nd International Meeting on Statistics 2026

Universidad Nacional de Colombia, Manizales Campus · Methods, Modeling and Applications

November 5–6, 2026 · La Nubia Campus, R and W blocks

Workshop and materials (ES · EN): `fastanalytics.co/encuentro-estadistica`

FastAnalytics S.A.S

El futuro habla en datos

¡Lo anticipamos con decisiones!

PATROCINADOR

2do Workshop Internacional de estadística

5 y 6 noviembre de 2026

UNIVERSIDAD NACIONAL DE COLOMBIA

El futuro habla en datos.

¡Lo anticipamos con decisiones!

FastAnalytics es una empresa de investigación, ciencia y tecnología que combina el modelado matemático, la analítica de datos y la inteligencia artificial para generar predicción estratégica.

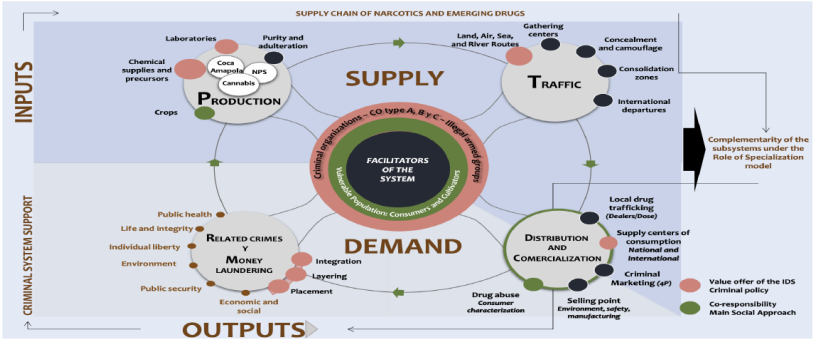
MODELADO MATEMÁTICO

ANALÍTICA DE DATOS · INTELIGENCIA ARTIFICIAL

fastanalytics.co



The illicit drug system (IDS): where the network lives



Illicit Drug System (IDS), built and translated from [Rodríguez, 2023].

Takeaway. The IDS shows that trafficking is one link in a system; our network is its logistics layer.

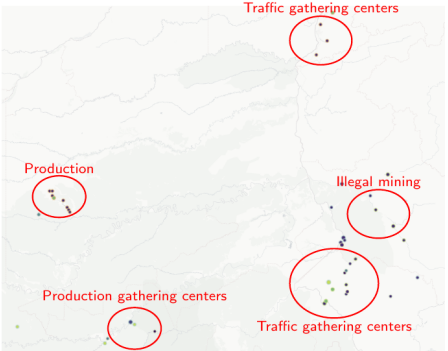
The study area: the Amazon shared by Colombia, Venezuela and Brazil





Takeaway. $\approx$ 706 thousand km^2 of jungle: 40 % larger than Spain, 12 % smaller than Turkey.

The territory speaks: high-clustering zones



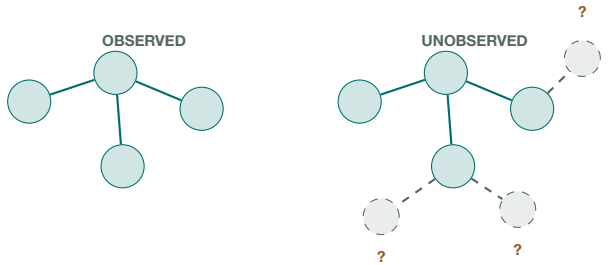
Takeaway. The overlap of high clustering and infrastructure is spatial evidence independent of the prediction.

The question is phrased over unobserved vertices and edges

Can we reconstruct a spatio-temporal criminal network by predicting its unobserved vertices and edges?

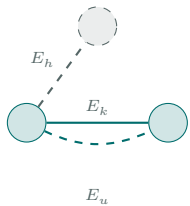
$G = (V, E)$ with $E \subseteq V \times V$ (directed: order matters)

Vertices: cells with communications · Edges: ordered pairs (u, v) of consecutive activity



The typology separates the observed from what will be inferred

Element	Notation	Definition
Known vertex	V_k	Location observed in intelligence
Known edge	E_k	Temporal link between known vertices
Unknown edge	E_u	Predicted between known vertices
Hidden edge	E_h	Involves at least one hidden vertex
Hidden vertex	V_h	Location inferred via hidden edges
Unknown vertex	V_u	Outside the study's geographic scope



solid: known
dashed teal: unknown
gray dashed: hidden

Takeaway. A hidden edge involves at least one hidden vertex; not necessarily two.

586 communications define the 2018 observable window

586	6	304	2018
HF records	groups A–F	active days	Jan 2 – Dec 31

OBSERVATION LIMITATIONS

approximate location · multiple speakers · mobile transmitters · overlapping territories · dual affiliations · unmonitored areas.

Takeaway. Every record contains time and location; its absence does not prove absence of activity.

The basic datum is an event ordered in space and time

$$e_a = (l_a, t_a), \qquad l_a = (x_a, y_a)$$
$$t_1 \leq t_2 \leq \cdots \leq t_{N_E}$$

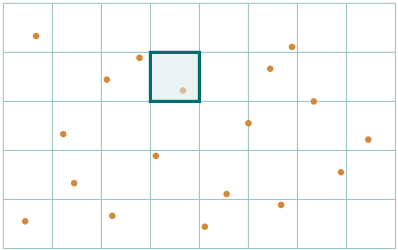


time orders events before connecting them

Each communication provides an **approximate location** and a **timestamp**. Chronological order precedes any connection.

Takeaway. First events are sorted; then aggregated into cells and windows.

Spatial resolution controls the aggregation of events



the **cell** (0.03°) defines the vertex

≈ 3.33 km · area ≈ 9.6 km^2

Cell	Min	Q1	Med.	Mean	Q3	Max
0.01°	1	1	1	1.24	1	12
0.02°	1	1	1	1.48	1	22
0.03°	1	1	1	1.76	2	25
0.04°	1	1	1	2.00	2	25
0.05°	1	1	1	2.20	2	39
0.06°	1	1	1	2.47	2	43

Chosen: 0.03° ≈ 3.33 km · area ≈ 9.6 km^2

Takeaway. The cell defines the vertex: coarser resolution aggregates more events but reduces spatial precision.

Connections occur between consecutive active days

$$I_k = [(k-1)\Delta t,\ k\Delta t)$$

$K = 304$ active days

$\Delta t = 1$ day



Takeaway. A jump joins consecutive active days, not necessarily consecutive calendar days.

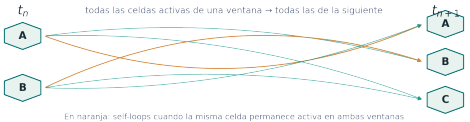
Complete connection between active cells of t_n and t_{n+1}

WEIGHT OF EACH ORDERED PAIR · EQUATION 1

$$w_{ij} = \sum_{k=1}^{K-1} n_i^{(k)}\, n_j^{(k+1)}$$

$n_i^{(k)}$: events of cell i on active day k.
$K = 304$: active days of 2018.
The product adds *co-activity*; the sum accumulates transitions.

IMAGE · COMPLETE RULE BETWEEN WINDOWS



every active origin connects to **all** active destinations; the same cell may form a self-loop (orange)

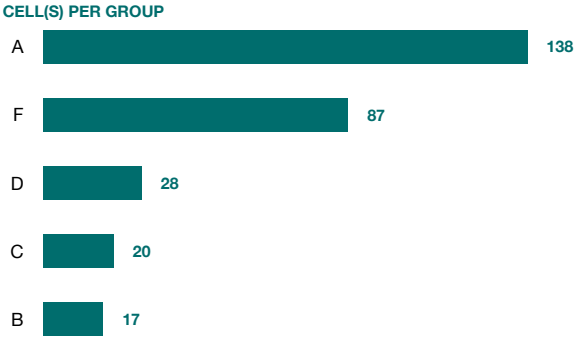
Takeaway. Every active origin connects to all active destinations; the same cell may form a self-loop.

The observed network contains 333 vertices and 1,105 edges

333
occupied cells (vertices)

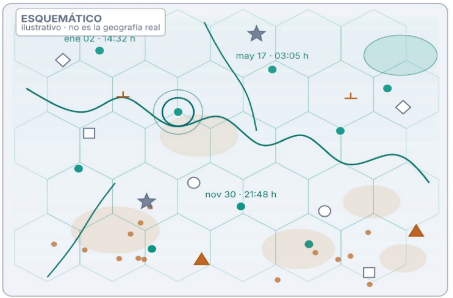
1,105
directed edges

Frequencies describe presence in cells; they are not by themselves a centrality measure.



Takeaway. The observed network is the reference for evaluating reconstruction and discovering missing elements.

Each vertex inherits its environmental context



«cada vértice hereda su contexto ambiental completo»

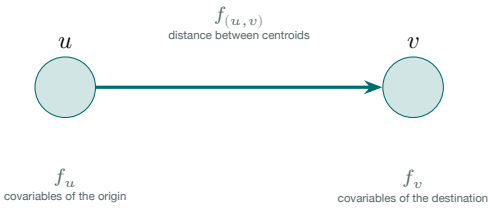
Takeaway. The bone is the HF communications; the flesh, the remote and climate layers; the contrast, OSM infrastructure to validate proximity.

Satellite and environmental data



Takeaway. Five raw sources (without the intercepted HF): what was seen from space and the physical environment of the vertex.

Each directed pair becomes one classification observation



$$P_{u,v} = \Psi(f_u, f_v, f_{(u,v)})$$

TARGET VARIABLE

$$G_{uv} \in \{0, 1\}$$

1 = observed edge
0 = pair without observed edge

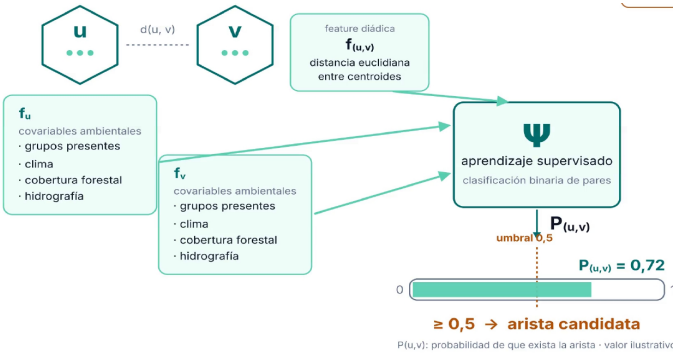
self-loops excluded in this dataset

IMBALANCED DATA

333 vertices $\Rightarrow 333 \times 332 \approx 111$ thousand directed pairs
1,105 observed edges $\Rightarrow \approx 1$ % positives

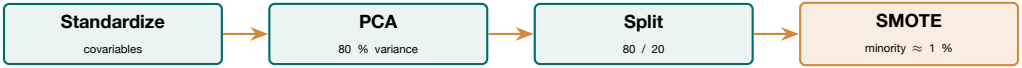
Takeaway. Reconstruction is cast as binary classification over pairs of cells.

Edge prediction = binary classification (scheme)



Takeaway. One row per pair (u, v): covariables of both vertices plus the centroid distance feed Ψ; the 0.5 threshold decides candidacy.

PCA and SMOTE: preprocessing under imbalance



PCA · PRINCIPAL COMPONENTS

SMOTE · SYNTHETIC MINORITY



$$x_{\text{new}} = x_i + \delta\,(x_{nn} - x_i), \quad \delta \in [0, 1]$$

Interpolates synthetic minority-class examples (edges) between close neighbors until training is balanced.

Takeaway. From raw vertex to balanced example: PCA compresses environmental information; SMOTE balances the minority class before training.

Specificity and logistic regression

SPECIFICITY · VALIDATION LOAD

$$\text{Specificity} = \frac{TN}{TN + FP}$$

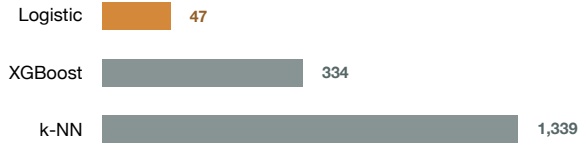
FP = potential candidate and verification workload; not a confirmed edge

LOGISTIC REGRESSION

$$\mathbf{p} = \frac{1}{1 + e^{-\mathbf{X}\beta}}$$

$\mathbf{X}$: principal components (80 % variance) · β: weights fitted by maximum likelihood

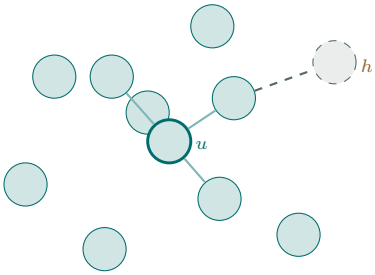
FALSE POSITIVES · TEST



bars on square-root scale

Takeaway. Fewer FP → a more manageable list for analysts; not more certainty about each candidate.

Homophily shrinks the search space of hidden vertices



$k = 3$ neighbors per vertex

hidden vertices share the environment (homophily)

333 known

environmental PCA

$k = 3$ **neighbors**

per known vertex

Candidate group

reduced search

500 seeds

selection stability

Assumption: hidden vertices share environmental characteristics with observed ones. It is a *homophily hypothesis*, not a proven property.

Takeaway. KNN restricts where to search; the 500 iterations measure selection persistence.

Prediction frequency measures stability across seeds

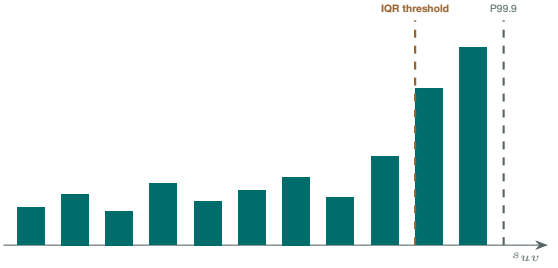
$$s_{uv} = \sum_{r=1}^{500} \mathbb{I}\left(p_{uv}^{(r)} > 0.5\right)$$

s_{uv}: positive classifications of pair (u, v) across 500 seeds (auxiliary presentation notation)

TWO CRITERIA

IQR
unknown edges

Percentile 99.9
hidden edges



selection frequency in 500 runs

Takeaway. A high frequency signals persistent selection; it is not a probability of existence.

Reconstruction preserves the main structure of the Chronnet

Metric	Observed	Reconstructed
Vertices	333	331
Edges	1,105	1,073
Density	0.00999	0.00982
Mean degree	6.64	6.48
Clustering	0.107	0.107
Diameter	16	16

97 %

edges recovered

99.4 %

vertices retained

The comparison evaluates structural preservation; the exponents are 2.65 and 2.70.

Takeaway. The reconstruction preserves the main structure of the observed Chronnet.

The IQR threshold selects 107 edges between known vertices

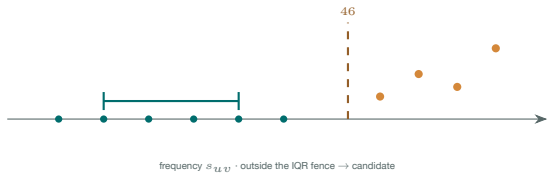
107

candidate unknown edges

$$s_{uv} > Q_3 + 1.5\,IQR = 46$$

the rule selects atypically high frequency values across the 500 runs

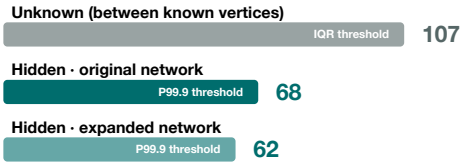
Both endpoints are **known vertices**.



Takeaway. The result is a prioritized list for verification, not 107 confirmed connections.

Two procedures converge on 84 unique hidden vertices

PREDICTED EDGES · APPLIED THRESHOLD



The 68 and 62 are **disjoint** (zero in common); the vertices they reveal **do overlap**: 19 shared.

COMPOSITION OF THE 84 HIDDEN VERTICES



35 unique (original) + 30 unique (expanded) + 19 shared = 84 hidden vertices to verify.

Takeaway. The 68 and 62 edges do not overlap; the vertices they reveal do: 35 + 30 + 19 shared = 84.

Critical infrastructure and processing chain

Dataset	Type	Source / resolution	Use in the model
Critical infrastructure	OSM vector	OpenStreetMap / Overpass Turbo · airfields, fuel, police, crops, quarries, docks	Spatial validation by proximity (5–30 km buffers)



Takeaway. Seven sources in total fuse into a single vector per vertex; OSM infrastructure also validates by proximity.

From a rank test to a territorial decision rule

HYPOTHESES ON MEDIANS

$$H_0 : \text{Median}(C_{\text{hidden}}) \geq \text{Median}(C_{\text{random}})$$
$$H_1 : \text{Median}(C_{\text{hidden}}) < \text{Median}(C_{\text{random}})$$

One-sided Wilcoxon rank test · 500 repetitions · buffers of 5, 10, 20 and 30 km.

$$R = \frac{1}{500} \sum_{r=1}^{500} \mathbb{I}(p_r \leq 0.05)$$

R: proportion of significant tests (auxiliary notation)

THE PAPER'S OPERATIONAL RULE



Application: quarries → attraction · docks → trade-off · always by scale and vertex group.

Takeaway. A mixed or inconclusive pattern does not imply irrelevance; it demands reading by infrastructure and scale.

Institutional validation · National Police

DIJIN · CRIMINAL INVESTIGATION AND INTERPOL DIRECTORATE

The **National Criminal Analysis Center (CENAC)** confirmed the **operational relevance** of the predicted hidden vertices and candidate edges: they coincide with zones and patterns known to police intelligence.

https://www.policia.gov.co/jefatura-nacional-del-servicio-de-policia/dijin

DIRAN · ANTI-NARCOTICS DIRECTORATE

The Anti-Narcotics Directorate validated the result in the Amazon context: the network's logistics layer is consistent with the trafficking systems the Police confront on the ground.

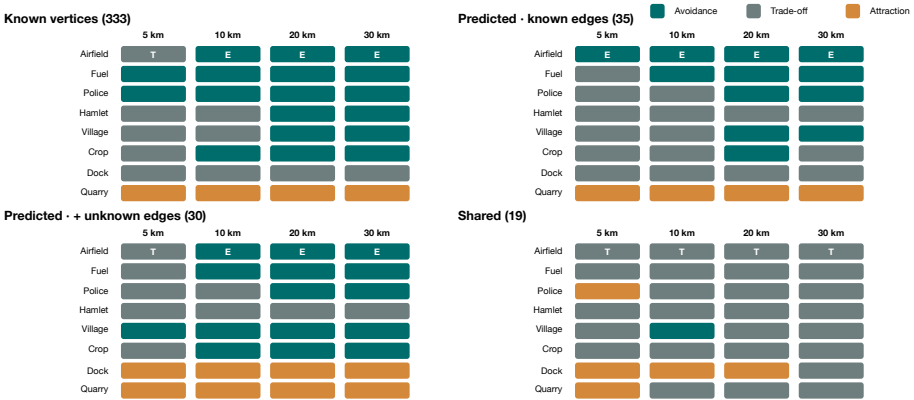
https://www.policia.gov.co/jefatura-nacional-del-servicio-de-policia/diran

What it validates and what it does not: the coincidence supports relevance and *orders field verification*; it does not confirm the physical presence of crime point by point.

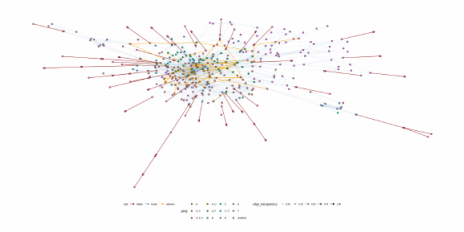
Takeaway. Operational validation turns predictions into prioritized hypotheses for police intelligence.

Three spatial logics — synthesis

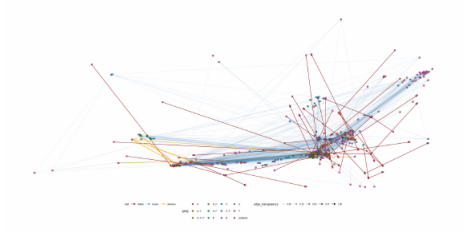
Avoidance: significant ($p \leq 0.05$) in ≥80 % of 500 runs · Attraction: ≤20 % and median $p \geq 0.8$ · Trade-off: the rest · Controls: 500 seeds · distinct thresholds



The reconstructed Chronnet, abstract and on the map



The complete reconstructed network, visualized by connection strength.



The same complete network, visualized by its real-world latitude and longitude.

Left: the full network by connection strength · Right: the same network at its real coordinates (lat-lon).

Thanks



Andrés Pérez-Coronado · **Speaker**

andres.perezc@fastanalytics.co

2nd International Meeting on Statistics 2026 · Universidad Nacional de Colombia, Manizales Campus

fcen.unal.edu.co/2wie2026

Workshop and materials (ES · EN): fastanalytics.co/encuentro-estadistica