Applied Mathematical Sciences, Vol. 20, 2026, no. 4, 231 - 247
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/ams.2026.919357

From Composite Indicators to DataOps Backbones: A Mathematical Architecture for Scalable SME Innovation Diagnosis

Cristina Di Letizia

Department of Social Science and Grant Office Service,
Third Mission and Major Projects Area
University of Foggia, Via Gramsci 89/91
71121 Foggia, Italy

Luca Grilli

Department of Economics
University of Foggia, Via Caggese, 1
71121 Foggia, Italy

Abstract

This paper extends the OECO_Sy model for SME innovation propensity from a determinant-level composite indicator to a scalable data architecture for university-led open innovation ecosystems. The starting point is the submitted companion paper by Di Letizia and Grilli, which provides a previous mathematical formalization of three coefficients, namely the Preliminary Innovation Propensity Coefficient (PIPC), the Management Innovation Propensity Coefficient (MIPC), and the Composite Coefficient of SME Innovation Propensity (CCSIP) [7]. The new contribution is the definition of a DataOps-oriented backbone, denoted by OECO_SyMDB, which makes the diagnostic process repeatable, traceable and extensible over time. We model the architecture as a family of source-indexed and time-indexed data objects, a versioned composition of transformations from landing to exploitation zones, and a deterministic Data Analysis Backbone that reproduces the OECO_Sy coefficients from the questionnaire sub-view only. This distinction corrects a common methodological ambiguity: the multi-source architecture may contain questionnaires, balance sheets, reports, innovation projects, industrial property records and briefs, but PIPC, MIPC and CCSIP must be computed only from the validated questionnaire determinants. Additional sources generate second-level indicators, not altered baseline coefficients. Several propositions establish boundedness, monotonicity, reproducibility under a fixed methodological version, and horizontal extensibility under schema invariance. A pilot proof-of-concept on the ten anonymized TCOREC firms is used only to verify deterministic recalculation of the coefficients, not to claim end-to-end validation of the complete five-source pipeline. The result is a mathematically controlled architecture for transforming a single-firm diagnostic model into a longitudinal observatory of SME innovation propensity.

Mathematics Subject Classification: 90B50, 62H20, 62P20, 68P15, 68T37

Keywords: Open innovation ecosystem, SME innovation, composite indicator, DataOps, data backbone, decision support, university Grant Office

1 Introduction

The OECO_Sy framework was originally introduced as a diagnostic model through which a university Grant Office can measure the innovation propensity of Small and Medium-sized Enterprises (SMEs) in the genesis phase of an open innovation ecosystem. Its first mathematical extension transformed three diagnostic questionnaires into twenty normalized determinants and three coefficients: PIPC, MIPC and CCSIP. This made the role of the university Expert measurable and contestable, rather than merely descriptive.

However, the first diagnostic layer raises a second mathematical and organizational problem. A coefficient computed once for one firm is useful for a pilot diagnosis, but it does not yet define an infrastructure. If the same diagnosis must be repeated over many firms, over different time instants, and under evolving methodological conventions, then the central question becomes one of reproducibility, lineage and scalability. The framework must preserve the logic of the composite indicators while allowing the system to ingest heterogeneous sources such as questionnaires, balance sheets, company reports, innovation projects, intellectual property titles and project briefs.

This paper addresses that problem. It develops a mathematical architecture, denoted by OECO_SyMDB, that translates OECO_Sy into a DataOps-oriented backbone. The contribution is not a generic theory of DataOps. It is a domain-specific adaptation of DataOps to a university-led diagnostic system for SME innovation. The architecture is formulated as a sequence of functions acting on source-indexed and time-indexed data objects. It distinguishes the

Data Management Backbone (DMB), which stores and governs multi-source data, from the deterministic Data Analysis Backbone (DAB), which computes the three OECO_Sy coefficients.

The distinction is essential. The complete data architecture may contain five source families: questionnaires, balance sheets, company reports, innovation projects and industrial property records, and project briefs. Yet the baseline coefficients PIPC, MIPC and CCSIP must not be recomputed from all these sources. They are defined by the validated questionnaire determinants. The other sources can enrich the diagnosis through second-level indicators, for example declared-activity coherence, economic-innovation performance, trajectory over time and brief-diagnosis coherence. They should not silently modify the original coefficients.

The paper also corrects two overclaims that may arise in moving from a pilot thesis chapter to a journal article. First, reproducing the manual coefficients from a CSV file built from the questionnaires is a verification of the deterministic DAB, not a full validation of the entire multi-source pipeline. Secondly, horizontal partitioning of the landing zone is a design condition for extensibility; it is not by itself an empirical proof of scalability in the operational sense of cost, latency, throughput and security. We therefore formulate scalability as a conditional property under schema invariance and bounded per-firm processing cost.

The paper is organized as follows. Section 2 summarizes the theoretical and methodological background. Section 3 recalls the OECO_Sy coefficient system. Section 4 formalizes the OECO_SyMDB data architecture. Section 5 states the main mathematical properties. Section 6 discusses verification, testing and governance. Section 7 reports the pilot proof-of-concept on the TCOREC dataset. Section 8 concludes with limitations and future research.

2 Background and research gap

Open innovation theory argues that firms should combine internal and external knowledge flows in order to innovate effectively [3]. Ecosystem theory extends this idea by emphasizing interdependence among firms, universities, public agencies, investors, users and complementary actors [1, 14]. In the genesis phase of an innovation ecosystem, roles are not yet stabilized. Dedehayir, Makinen and Ortt [5] classify ecosystem roles and identify the university as a possible Expert, that is, an actor supporting value creation by providing knowledge, coordination, mentoring and access to resources.

For SMEs, the role of an Expert is especially relevant because internal resources are limited and innovation capabilities are often uneven. The Oslo Manual interprets innovation capability as a multidimensional construct involving managerial, human-resource and technological dimensions [20]. The problem is that such dimensions are rarely available as standardized firm-level data at the beginning of university-enterprise collaboration. The organizational framework was introduced by Di Letizia, Grilli and Di Noia [6], while its determinant-level composite-indicator formalization was developed in the submitted companion paper by Di Letizia and Grilli [7]. Its purpose is not to replace managerial judgement, but to make the first diagnostic judgement explicit.

The submitted OECO_Sy composite-indicator paper [7] solves the measurement problem at the level of one diagnostic run. The present paper solves the persistence problem: how to store, recompute and extend that diagnosis. This problem belongs to the intersection of three literatures. The first is composite-indicator methodology, which requires explicit normalization, weighting, aggregation and uncertainty analysis [19]. The second is DataOps, which treats data pipelines as versioned, tested, governed and continuously improved processes [18, 22, 27]. The third is Design Science Research, which is appropriate when the research output is an artefact, such as a model, method or architecture [12, 21].

DataOps is particularly useful here because OECO_Sy is not a high-volume big-data system at inception. Its critical dimension is not initially volume, but traceability. Each diagnostic coefficient depends on questionnaire versions, coding rubrics, missing-value rules, branching rules, weights, thresholds and timestamps. If these objects are not versioned, old coefficients become ambiguous when the method is revised. If they are versioned, the system can grow into a territorial observatory without losing reproducibility.

The research gap is therefore the absence of a mathematically explicit bridge between a composite indicator of SME innovation propensity and the data architecture required to repeat it over time and across firms. The paper fills this gap by specifying (i) a versioned scoring function, (ii) a DMB mapping heterogeneous sources into an exploitation view, (iii) a deterministic DAB for coefficient computation, (iv) a lineage object for reproducibility and (v) a conditional scalability property.

3 The OECO_Sy coefficient layer

Let $\mathcal{I} = \{1, \dots, n\}$ be a finite set of firms. The OECO_Sy questionnaire determinants are partitioned into three sets

$$D_1, \qquad D_2, \qquad D_3,$$

where D_1 contains the eight determinants of Q.1, D_2 the five determinants of Q.2, and D_3 the seven determinants of Q.3. Thus $|D_1| = 8$, $|D_2| = 5$, $|D_3| = 7$ and $|D_1 \cup D_2 \cup D_3| = 20$.

For each firm i and determinant j, let $x_{ij} \in [0, 1]$ be the normalized score. Some determinants may be structurally unavailable because of branching. Therefore, for $k = 1, 2, 3$, let

$$A_{k,i} \subseteq D_k$$

be the set of determinants effectively available for firm i in questionnaire block k.

Definition 3.1 *The block score of firm i in questionnaire block k is*

$$Q_{k,i} = \frac{1}{|A_{k,i}|} \sum_{j \in A_{k,i}} x_{ij}, \qquad k = 1, 2, 3. \tag{1}$$

The Preliminary Innovation Propensity Coefficient is defined at determinant level, not as a simple mean of the two block means:

$$PIPC_i = \frac{1}{|A_{1,i}| + |A_{2,i}|} \left(\sum_{j \in A_{1,i}} x_{ij} + \sum_{j \in A_{2,i}} x_{ij} \right). \tag{2}$$

The Management Innovation Propensity Coefficient is

$$MIPC_i = \frac{1}{|A_{3,i}|} \sum_{j \in A_{3,i}} x_{ij}. \tag{3}$$

Finally,

$$CCSIP_i = \frac{PIPC_i + MIPC_i}{2}. \tag{4}$$

The diagnostic gap

$$G_i = MIPC_i - PIPC_i \tag{5}$$

indicates whether the internal organizational substrate of the firm is stronger or weaker than the formalized side of innovation measured through funding readiness, innovation history, artificial intelligence adoption and industrial property maturity.

Equal aggregation is not interpreted as a claim that all determinants have equal causal importance. It is a transparent baseline suitable for a pilot model in which no external validation criterion is yet available. Future versions may introduce expert-calibrated or data-driven weights, provided that each version is stored and old scores remain reproducible.

4 The OECO_SyMDB architecture

4.1 Data objects and sources

Let $\mathcal{T}$ be a discrete set of observation times and let

$$\mathcal{S} = \{Q, B, R, P, H\}$$

be the set of source families, where Q denotes OECO_Sy questionnaires, B balance-sheet data, R company reports, P innovation projects and industrial property records, and H project briefs. For each $s \in \mathcal{S}$, firm $i \in \mathcal{I}$ and time $t \in \mathcal{T}$, let

$$D_{s,i,t} \in \mathcal{D}_s \cup \{\perp\}$$

be the raw data object supplied by source s, where $\perp$ denotes absence of data.

The five source families have different epistemic status. The source Q is the only source used to compute the baseline coefficients. The sources B, R, P and H are complementary sources used to construct second-level indicators. In particular, H does not enter PIPC, MIPC or CCSIP; it supports brief-diagnosis coherence in the Formation State.

Definition 4.1 *For each firm i and time t, the raw multi-source vector is*

$$D_{i,t} = (D_{Q,i,t}, D_{B,i,t}, D_{R,i,t}, D_{P,i,t}, D_{H,i,t})\,. \tag{6}$$

The questionnaire sub-vector is $D^{Q}_{i,t} = D_{Q,i,t}$.

4.2 Versioned transformations

Let θ_v denote the methodological version used at run v. It includes weights, thresholds, coding rubrics, missing-value rules, branching rules, schema versions and code versions:

$$\theta_v = (w_v, \tau_v, r_v, m_v, b_v, \sigma_v, c_v). \tag{7}$$

The Data Management Backbone is represented by four transformations:

$$\phi_L, \quad \phi_F, \quad \phi_T, \quad \phi_E,$$

corresponding to landing, formatted, trusted and exploitation zones. They act on data objects and metadata. Since methodology may evolve, the transformations are versioned.

Definition 4.2 *The OECO_Sy Data Management Backbone under version v is the composition*

$$\Phi_v = \phi_{E,v} \circ \phi_{T,v} \circ \phi_{F,v} \circ \phi_{L,v}. \tag{8}$$

For each firm-time pair (i, t),

$$E^{(v)}_{i,t} = \Phi_v(D_{i,t}) \tag{9}$$

is the exploitation view generated from all available sources.

The trusted-zone transformation is itself decomposed as

$$\phi_{T,v} = \rho_{out,v} \circ \rho_{clean,v} \circ \rho_{dup,v}, \tag{10}$$

where ρ_{dup} removes duplicates, ρ_{clean} handles missing or inconsistent values, and ρ_{out} identifies or flags outliers. The order is part of the model. Outlier detection before duplicate handling may distort the empirical distribution from which outlier thresholds are computed.

The exploitation view is decomposed into a questionnaire sub-view and complementary sub-views:

$$E^{(v)}_{i,t} = E^{(v)}_{Q,i,t} \cup E^{(v)}_{B,i,t} \cup E^{(v)}_{R,i,t} \cup E^{(v)}_{P,i,t} \cup E^{(v)}_{H,i,t}. \tag{11}$$

Only $E^{(v)}_{Q,i,t}$ is an admissible input for the deterministic coefficient computation.

4.3 Deterministic DAB

Definition 4.3 *The deterministic Data Analysis Backbone for OECO_Sy coefficients is the function*

$$F_v : E^{(v)}_{Q,i,t} \longrightarrow \left(PIPC^{(v)}_{i,t}, MIPC^{(v)}_{i,t}, CCSIP^{(v)}_{i,t}, G^{(v)}_{i,t}\right), \tag{12}$$

where the four outputs are computed by equations (2)-(5) using the availability sets and aggregation rules encoded in θ_v.

The composition $F_v \circ \Phi_v$ should be understood with a projection. Formally, if

$$\pi_Q(E^{(v)}_{i,t}) = E^{(v)}_{Q,i,t},$$

then

$$(F_v \circ \pi_Q \circ \Phi_v)(D_{i,t}) = \left(PIPC^{(v)}_{i,t}, MIPC^{(v)}_{i,t}, CCSIP^{(v)}_{i,t}, G^{(v)}_{i,t}\right). \tag{13}$$

This is the coefficient vector for firm i at time t under version v. This notation prevents the methodological error of treating balance sheets, reports or briefs as hidden inputs to PIPC, MIPC or CCSIP.

5 Mathematical properties

Theorem 5.1 (Boundedness and monotonicity) *Assume $x_{ij} \in [0,1]$ for all available determinants and $|A_{k,i}| > 0$ for $k = 1,2,3$. Then*

$$PIPC_i, MIPC_i, CCSIP_i \in [0,1].$$

Moreover, if the availability sets are fixed and one or more available determinant scores weakly increase while all others remain unchanged, then the corresponding coefficient affected by those determinants weakly increases.

Proof. Each block score and each coefficient is a convex combination of numbers in $[0,1]$. Therefore it also belongs to $[0,1]$. Monotonicity follows because all coefficients in the convex combinations are non-negative. □

Theorem 5.2 (Reproducibility equivalence) *Fix a methodological version v and a firm-time pair (i,t). Suppose that the manual OECO_Sy calculation and the deterministic DAB use the same normalized determinant matrix, the same availability sets, and the same version parameters θ_v. Then the manual coefficient vector and the DAB coefficient vector are identical in exact arithmetic. Under floating-point arithmetic, the observed error is bounded by a declared numerical tolerance ε_v.*

Proof. Under the stated assumptions both computations apply equations (2)-(5) to the same finite set of input numbers and the same denominators. Hence the algebraic expressions are identical. Floating-point computation may introduce rounding error; this is not a methodological difference but a numerical convention, and it is controlled by ε_v. □

Theorem 5.3 (Horizontal extensibility under schema invariance) *Let $\mathcal{I}_n = \{1,\dots,n\}$ and assume that the source schemas, canonical mappings and quality rules are invariant across firms. If the landing zone is partitioned by firm,*

$$D_L = \bigcup_{i\in\mathcal{I}_n} D_{L,i}, \qquad D_{L,i} \cap D_{L,h} = \emptyset \quad (i \neq h),$$

and $\phi_{F,v}$, $\phi_{T,v}$ and $\phi_{E,v}$ are firm-invariant, then

$$\Phi_v(D_L) = \bigcup_{i\in\mathcal{I}_n} \Phi_v(D_{L,i}). \tag{14}$$

Consequently, adding a new firm does not require redesign of the formatted, trusted or exploitation transformations, provided that the new firm conforms to the same schemas and governance rules.

Proof. By assumption the transformations after the landing zone are defined by source type and schema, not by firm identity. Since the landing partitions are disjoint, applying the same transformations to the union is equivalent to applying them to each partition and then taking the union of the resulting exploitation views. □

The theorem is deliberately conditional. It proves architectural reusability, not empirical scalability in the full operational sense. Let

$$C(n) = C_0 + \sum_{i=1}^{n} c_i + R(n) \tag{15}$$

be the total cost of processing n firms, where C_0 is the fixed design cost, c_i is the marginal ingestion and verification cost of firm i, and $R(n)$ is a resource-congestion term. Horizontal extensibility implies that no redesign term is needed for each new firm. It does not imply $R(n) = 0$. Load testing is required to estimate latency, throughput, error rates and security overhead.

6 Second-level indicators

The exploitation view contains more information than the questionnaire subview. However, this additional information should create second-level indicators rather than alter the validated coefficients. Let

$$Y_{i,t} = F_v(E^{(v)}_{Q,i,t})$$

be the coefficient vector. Let $z_{P,i,t}$ denote a normalized documentary innovation-activity signal derived from projects and IP records; let $z_{B,i,t}$ denote a vector of economic performance variables; let $z_{H,i,t}$ denote an ordinal brief-maturity or brief-coherence variable; and let $z_{R,i,t}$ denote report metadata or future text features.

Four families of indicators are then admissible:

$$C^{decl}_{i,t} = 1 - |CCSIP^{(v)}_{i,t} - z_{P,i,t}|, \tag{16}$$

$$C^{perf}_{i,t,\Delta} = g(z_{B,i,t+\Delta}, Y_{i,t}), \tag{17}$$

$$C^{traj}_{i,t,\Delta} = CCSIP^{(v')}_{i,t+\Delta} - CCSIP^{(v)}_{i,t}, \tag{18}$$

$$C^{brief}_{i,t} = h(z_{H,i,t}, Y_{i,t}). \tag{19}$$

Here g and h must be specified by the analyst. For instance, g may be a correlation, regression or classification model when sufficient longitudinal data become available. The function h may be an ordinal coding rule measuring whether the brief is coherent with, more mature than, or disconnected from the diagnostic profile.

This separation has two advantages. First, the baseline diagnostic scale remains stable and reproducible. Secondly, future observatory-level analytics can be developed without rewriting the meaning of PIPC, MIPC and CCSIP.

7 Lineage, testing and governance

For each execution of the deterministic DAB, the architecture must store a lineage object

$$\begin{aligned} L_{i,t}^{(v)} = (&source\_id, timestamp, operator\_id, rubric\_version, \\ &weights\_version, threshold\_version, schema\_version, \\ &code\_hash, input\_hash, output\_hash). \end{aligned} \tag{20}$$

This is not a bureaucratic addition. It is the condition that makes historical coefficients meaningful after methodological updates.

Test	Acceptance rule
Schema test	Every input table contains firm identifier, source, determinant or variable name, value, timestamp and schema version.
Range test	Raw questionnaire scores belong to the declared input range and normalized scores belong to $[0,1]$.
Branching test	Excluded determinants have an explicit missing type and do not enter the denominator of the coefficient.
Recalculation test	A known mini-dataset is recomputed with maximum absolute error not exceeding ε_v, for example 10^{-6} in an internal implementation or 10^{-3} under rounded reporting.
Duplicate test	Duplicate records with the same firm, source, determinant and time are flagged before cleaning and outlier detection.
Lineage test	Each output record stores code, schema, weight, threshold and rubric versions.
Access-control test	Sensitive questionnaires, reports and briefs are visible only to authorized roles; aggregated observatory outputs satisfy the declared anonymity rule.

Table 1: Minimum test suite for a reproducible OECO_SyMDB implementation.

Governance is not external to the mathematics of the model. If the version object is missing, the coefficient vector becomes ambiguous. If the branching flag is missing, equations (1)-(3) may use the wrong denominator. If access control is missing, confidential innovation information cannot be ethically used in an observatory. The governance layer must therefore define roles such as data owner, scientific coordinator, data steward, data engineer, analyst, Data Protection Officer and external provider. For public sources, such as balance sheets or industrial property databases, service-level agreements must specify access frequency, reuse limits and update rules. For sensitive sources, non-disclosure agreements and privacy-by-design policies must regulate processing.

8 Pilot proof-of-concept

The TCOREC pilot dataset contains ten anonymized firms already diagnosed by OECO_Sy. The submitted companion paper [7] reports PIPC, MIPC and CCSIP for those firms and interprets the empirical pattern. The purpose here is different. The pilot is used to check whether the deterministic DAB can reproduce the manual coefficient calculation from a machine-readable input table.

The relevant input table contains one row for each firm-determinant pair. Its minimum columns are

$$(pmi\_id, source, determinant, value, timestamp, schema\_version).$$

Values are ingested in a raw range, rescaled to $[0, 1]$, grouped by firm and questionnaire block, and aggregated through equations (1)-(4). Missing determinants caused by branching are omitted from the denominator and recorded through a missing-type flag.

Algorithm 1. Deterministic DAB for OECO_Sy coefficients.

1. Read the input table and verify the schema.
2. For every row, rescale the score to $[0, 1]$ and verify the range.
3. For every firm i, build $A_{1,i}$, $A_{2,i}$ and $A_{3,i}$ from the branching flags.
4. Compute $PIPC_i$, $MIPC_i$, $CCSIP_i$ and G_i using equations (2)-(5).
5. Attach the lineage object $L_i^{(v)}$ to the output.
6. Compare the output with the manual reference values and report the maximum absolute deviation.

In the TCOREC proof-of-concept, the recalculation reproduces the manual coefficient table, with only a rounding deviation of order 10^{-3} in one reported value due to different rounding conventions between the spreadsheet and the script. This is consistent with Theorem 5.2. It verifies that the deterministic aggregation layer can be transferred to a pipeline.

The result must be interpreted carefully. It does not prove that the complete OECO_SyMDB has already integrated all five source families. It verifies one necessary component: the deterministic DAB applied to the questionnaire sub-view. The end-to-end pipeline, including balance sheets, reports, projects, IP records and briefs, requires a prospective pilot with real source acquisition, data-quality tests and governance procedures.

Firm	PIPC	MIPC	CCSIP
IND-01	0.799	0.726	0.763
AGR-03	0.660	0.664	0.662
TUR-01	0.517	0.773	0.645
AGR-02	0.672	0.607	0.640
HEA-01	0.494	0.702	0.598
IND-02	0.519	0.660	0.590
RET-01	0.478	0.695	0.587
ENE-01	0.481	0.681	0.581
HEA-02	0.405	0.683	0.544
AGR-01	0.334	0.672	0.503

Table 2: OECO_Sy coefficients used as the manual reference in the deterministic DAB proof-of-concept.

The table also illustrates why the observatory perspective is meaningful. In most firms MIPC exceeds PIPC, suggesting that organizational conditions may be stronger than formalized innovation assets. Nevertheless, such a pattern is only pilot evidence. With $N = 10$ it should not be generalized to the whole regional production system. Its proper use is to identify a diagnostic hypothesis to be tested when the number of firms and time points increases.

9 Toward a territorial observatory

The OECO_SyMDB architecture provides the technical precondition for a territorial observatory, but the observatory should be developed in stages. The first stage is descriptive: dashboards, distributions of PIPC, MIPC and CCSIP, determinant-level heat maps and sector comparisons. The second stage is unsupervised: clustering and dimensionality reduction, where the aim is to identify firm profiles rather than predict outcomes. The third stage is text-based: natural language processing of reports and briefs, after a validated coding protocol and sufficient text volume exist. The fourth stage is supervised prediction: models for grant success, IP filing, revenue growth or transition from diagnosis to formal open-innovation agreement. This last stage requires longitudinal outcomes and must be evaluated through training, validation and test splits.

Let $O_{i,t+\Delta}$ denote an external outcome observed after time lag Δ. Examples are a successful grant application, a new IP title, a signed agreement or an increase in an innovation-related economic indicator. A predictive model is admissible only when the dataset

$$\{(Y_{i,t}, z_{B,i,t}, z_{P,i,t}, z_{H,i,t}, O_{i,t+\Delta})\}_{i,t}$$

contains enough observations to estimate and validate the mapping

$$\widehat{O}_{i,t+\Delta} = M(Y_{i,t}, z_{B,i,t}, z_{P,i,t}, z_{H,i,t}). \tag{21}$$

Until then, machine learning remains a roadmap, not an empirical result.

At the aggregated level, privacy constraints are decisive. For a territory-sector cell C, let n_C be the number of firms in the cell. An observatory rule should suppress or coarsen any output for which

$$n_C < k, \tag{22}$$

where k is a declared minimum cell size. This is a simple form of statistical disclosure control. For sensitive indicators, additional perturbation or differential privacy can be considered. The goal is to produce territorial knowledge without exposing individual business strategies.

10 Conclusions

This paper has extended the OECO_Sy composite-indicator model into a mathematical data architecture for repeatable and scalable SME innovation diagnosis. The main contribution is the OECO_SyMDB backbone, defined as a versioned composition of transformations from source acquisition to exploitation view, together with a deterministic Data Analysis Backbone for computing PIPC, MIPC and CCSIP from the questionnaire sub-view.

The paper contributes to applied mathematics in three ways. First, it formalizes a university-led innovation support process as a family of source-indexed, time-indexed and versioned mappings. Secondly, it establishes boundedness, monotonicity, reproducibility and conditional horizontal extensibility of the diagnostic architecture. Thirdly, it separates baseline coefficients from second-level indicators, thereby preserving the semantics of the original composite indicator while enabling multi-source enrichment.

The TCOREC proof-of-concept confirms deterministic recalculation of the coefficients, but it does not validate the full five-source pipeline. That distinction is methodological rather than cosmetic. A rigorous journal article must not present an architecture as empirically validated merely because one of its components has been reproduced on a pilot dataset. The next research step is therefore a prospective pilot including real ingestion of balance sheets, reports, projects, IP records and briefs, a minimum automated test suite, a canonical data model, and explicit governance rules.

Future research should extend the sample to a longitudinal panel, validate the Q.1 coding rubric through inter-rater reliability, calibrate weights through expert and data-driven procedures, test second-level indicators against external outcomes, and evaluate scalability through measurable latency, throughput, error-rate and cost indicators. Under these conditions, OECO_Sy can evolve from a diagnostic coefficient system into a territorial observatory of SME innovation propensity.

Acknowledgements. The authors acknowledge the Grant Office Service of the University of Foggia and the firms participating in the TCOREC pilot case. The SME-level information used in the proof-of-concept is anonymized and is not publicly disclosed because it contains confidential innovation and intellectual-property information.

References

[1] R. Adner, Match your innovation strategy to your innovation ecosystem, *Harvard Business Review*, **84** (2006), no. 4, 98 - 107.

[2] S. Azzabi, Z. Alfughi and A. Ouda, Data lakes: a survey of concepts and architectures, *Computers*, **13** (2024), no. 7, 183. https://doi.org/10.3390/computers13070183

[3] H.W. Chesbrough, *Open Innovation: The New Imperative for Creating and Profiting from Technology*, Harvard Business School Press, Boston, 2003.

[4] X. Coll Ribas, *A DataOps Reference Architecture for Data Science*, Bachelor thesis, Universitat Politecnica de Catalunya, Facultat d'Informatica de Barcelona, 2023.

[5] O. Dedehayir, S.J. Makinen and R. Ortt, Roles during innovation ecosystem genesis: a literature review, *Technological Forecasting and Social Change*, **136** (2018), 18 - 29. https://doi.org/10.1016/j.techfore.2016.11.028

[6] C. Di Letizia, L. Grilli and A.E. Di Noia, Innovation ecosystem and university system: the framework proposal for SMEs by the Grant Office Service of the University of Foggia, in *Quality Innovation and Sustainability*, Editors: Federica Murmura, Laura Bravi, Giada Pierli, Fabio Musso,

João Carlos Gonçalves dos Reis, Gilberto Santos, Springer Cham, ISBN: 978-3-032-19627-9, Series ISSN 2198-7246, 2026.

[7] Cristina Di Letizia, Luca Grilli, From framework to diagnostic composite indicator: the OECO_Sy model for SME innovation propensity in university-led open innovation ecosystems, *Applied Mathematical Sciences*, **20**, 2026, no. 4, 213-230.
https://doi.org/10.12988/ams.2026.919341

[8] EPO and EUIPO, *Intellectual Property Rights and Firm Performance in the European Union: Firm-level Analysis Report*, European Patent Office and European Union Intellectual Property Office, January 2025.

[9] J. Ereth and W.W. Eckerson, DataOps: industrializing data and analytics, Eckerson Group Research Report, 2018.

[10] European Commission, *Italy 2025 Digital Decade Country Report*, Directorate-General for Communications Networks, Content and Technology, 2025.

[11] V. Garousi, M. Felderer and M.V. Mantyla, Guidelines for including grey literature and conducting multivocal literature reviews in software engineering, *Information and Software Technology*, **106** (2019), 101 - 121.
https://doi.org/10.1016/j.infsof.2018.09.006

[12] A.R. Hevner, S.T. March, J. Park and S. Ram, Design science in information systems research, *MIS Quarterly*, **28** (2004), no. 1, 75 - 105.
https://doi.org/10.2307/25148625

[13] M. Iansiti and R. Levien, *The Keystone Advantage: What the New Dynamics of Business Ecosystems Mean for Strategy, Innovation, and Sustainability*, Harvard Business School Press, Boston, 2004.

[14] M.G. Jacobides, C. Cennamo and A. Gawer, Towards a theory of ecosystems, *Strategic Management Journal*, **39** (2018), no. 8, 2255 - 2276.
https://doi.org/10.1002/smj.2904

[15] T. Justy, E. Pellegrin-Boucher, D. Lescop, J. Granata and S. Gupta, On the edge of big data: drivers and barriers to data analytics adoption in SMEs, *Technovation*, **127** (2023), 102850.
https://doi.org/10.1016/j.technovation.2023.102850

[16] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed., Sage Publications, Thousand Oaks, 2004.
https://doi.org/10.4135/9781071878781

[17] P. Mayring, *Qualitative Content Analysis: Theoretical Foundation, Basic Procedure and Software Solution*, GESIS - Leibniz-Institut fur Sozialwissenschaften, Klagenfurt, 2014. https://doi.org/10.4135/9781036231798

[18] A.R. Munappy, J. Bosch and H.H. Olsson, Data pipeline management in practice: challenges and opportunities, in *Proceedings of the 46th Euromicro Conference on Software Engineering and Advanced Applications*, IEEE, 2020, 168 - 175. https://doi.org/10.1109/seaa.2019.00030

[19] OECD and European Commission Joint Research Centre, *Handbook on Constructing Composite Indicators: Methodology and User Guide*, OECD Publishing, Paris, 2008. https://doi.org/10.1787/9789264043466-en

[20] OECD and Eurostat, *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation*, 4th ed., OECD Publishing, Paris, 2018. https://doi.org/10.1787/9789264304604-en

[21] K. Peffers, T. Tuunanen, M.A. Rothenberger and S. Chatterjee, A design science research methodology for information systems research, *Journal of Management Information Systems*, **24** (2007), no. 3, 45 - 77. https://doi.org/10.2753/MIS0742-1222240302

[22] M.A. Rodriguez, R. Munoz, V. Naranjo and D. Ortiz, DataOps for data analytics: a systematic mapping study, *IEEE Access*, **8** (2020), 216971 - 216987.

[23] M. Saunila, Performance measurement approach for innovation capability in SMEs, *International Journal of Productivity and Performance Management*, **65** (2016), no. 2, 162 - 176. https://doi.org/10.1108/IJPPM-08-2014-0123

[24] Y.L. Simmhan, B. Plale and D. Gannon, A survey of data provenance in e-science, *SIGMOD Record*, **34** (2005), no. 3, 31 - 36. https://doi.org/10.1145/1084805.1084812

[25] M.L. Tushman and D.A. Nadler, Information processing as an integrating concept in organizational design, *Academy of Management Review*, **3** (1978), no. 3, 613 - 624. https://doi.org/10.5465/amr.1978.4305791

[26] S.L. Vargo, P.P. Maglio and M.A. Akaka, On value and value co-creation: a service systems and service logic perspective, *European Management Journal*, **26** (2008), no. 3, 145 - 152. https://doi.org/10.1016/j.emj.2008.04.003

[27] J. Xu, H. Naseer, S. Maynard and J. Filippou, Using analytical information for digital business transformation through DataOps: a review and conceptual framework, *Australasian Journal of Information Systems*, **28** (2024), 1 - 39. https://doi.org/10.3127/ajis.v28.4215

Received: March 15, 2026; Accepted: July 17, 2026
Published: July 23, 2026