TMF Workshop: Anonymization tools and their practical relevance (for biomedical research)

ARX

A Comprehensive Tool for Anonymizing Biomedical Data

Fabian Prasser, Florian Kohlmayer, Klaus A. Kuhn

Lehrstuhl für Medizinische Informatik
Institut für Medizinische Statistik und Epidemiologie
Klinikum rechts der Isar der TU München

ARX: Towards useful data anonymization

- **Usability has many dimensions**
  - Ability to balance data utility with privacy requirements
- **Need to support a broad spectrum of methods**
  - Privacy models
  - Transformation models
  - Methods for analyzing data utility
  - Methods for analyzing risks
- **Further “non-functional” requirements**
  - Integrated and harmonized: ARX is not a “tool box”
  - Compatibility (syntactic and semantic)
  - Performance and scalability
  - Intuitive visualization and parameterization
  - Provide methods to end-users as well as programmers

ARX: Highlights

- **Compatibility**
  - **Built-in data import facilities**
    - Relational databases (MS SQL, DB2, SQLite, MySQL)
    - MS Excel
    - CSV (all common formats, auto-detection)
  - **Support for different data types and scales of measure**
    - Strings (with nominal and ordinal scale)
    - Dates (interval scale)
    - Numbers (ratio scale)
    - Automatic detection of data types and formats
  - **Methods for handling and cleaning low-quality data**
    - Handles missing and invalid values correctly
    - In privacy models, transformation methods, visualizations
    - Manual removal of tuples, query interface, find & replace

ARX: Highlights

- **Flexible transformation methods**
  - **Global recoding**
    - Full-domain generalization
    - Top- & bottom coding
  - **Local recoding**
    - Tuple suppression
  - **Fully integrated and parameterizable**
    - Importance of attributes, suitability of methods
- **Functional representations of transformation rules**
  - Especially functional representations of hierarchies
  - Support for categorical and continuous variables (categorization)
- **Multiple methods for measuring data utility**
  - Parametrizable, e.g., with different aggregate functions
  - Use functional representations of transformation rules

ARX: Highlights

- **Multiple apriori privacy models**
  - k-Anonymity
  - ℓ-Diversity (distinct, entropy, recursive-(c, ℓ))
  - t-Closeness (equal and hierarchical ground distance)
  - δ-Presence
- **Multiple methods for risk-based anonymization**
  - Sample characteristics
    - Average cell size
    - Sample uniqueness
  - Super-population models
    - Decision rule by Dankar et al.
    - Based on models by Pitman, Zayatz and the SNB model
- **Support for arbitrary combinations of these models**
  - Optimal solution within our coding model

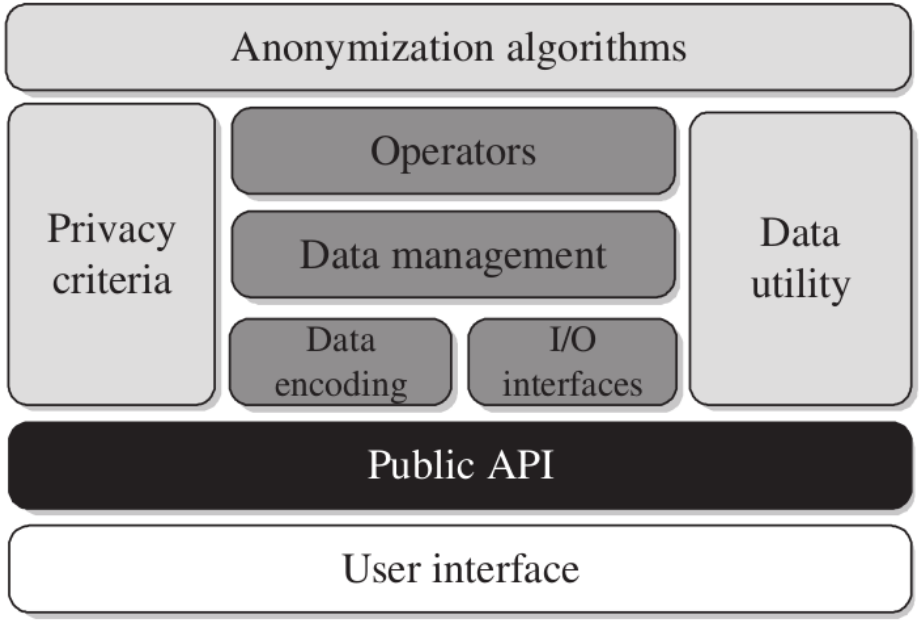
ARX: Highlights

- **Scalability:** ARX can handle large datasets (several million data entries) on commodity hardware
- **Efficient in-memory data management engine**
  - Works with compressed data representations
  - Tight coupling between transformation operators and the „database kernel”
  - Provides a space-time trade-off
- **Optimized search strategy:** Based on multiple pruning strategies
- **Efficient implementations of further complex tasks**
  - Evaluations of privacy criteria (e.g. t-closeness, δ-presence)
  - Methods for solving non-linear equation systems
  - Background jobs in the user interface

ARX: Highlights

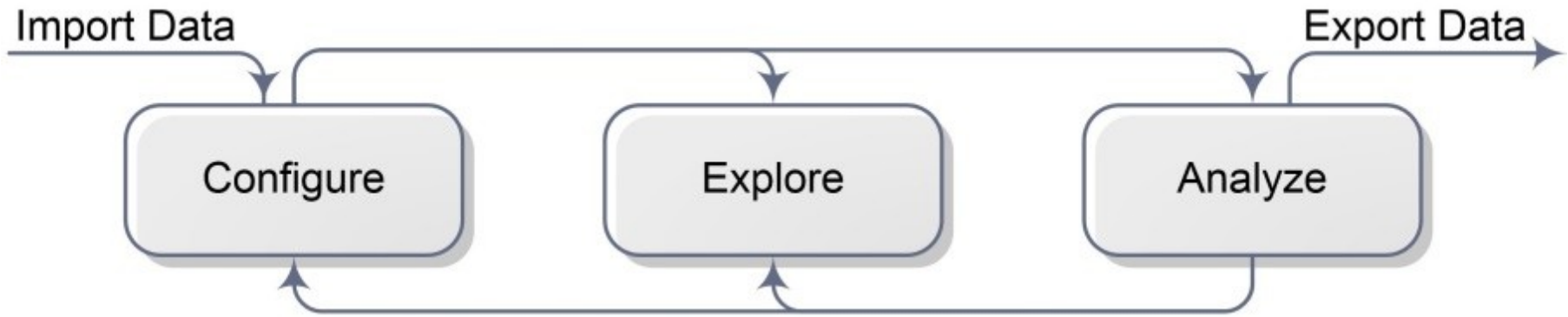
- **Comprehensive graphical interface**
  - Scalablity comparable to the ARX API
  - Supports all methods provided by the ARX API
  - Cross-platform (Windows, Linux/GTK, OSX) with native interfaces
  - Available as binary distributions with installers

- **Independent API**
  - User interface sits on top of the API
  - Java library
  - All methods provided by ARX are first-class citizens in both worlds



ARX: Anonymization workflow

- Iterative process to successively refine transformations
- Supported by the scalability of our framework
- Three (repeating) steps, mapped to four perspectives

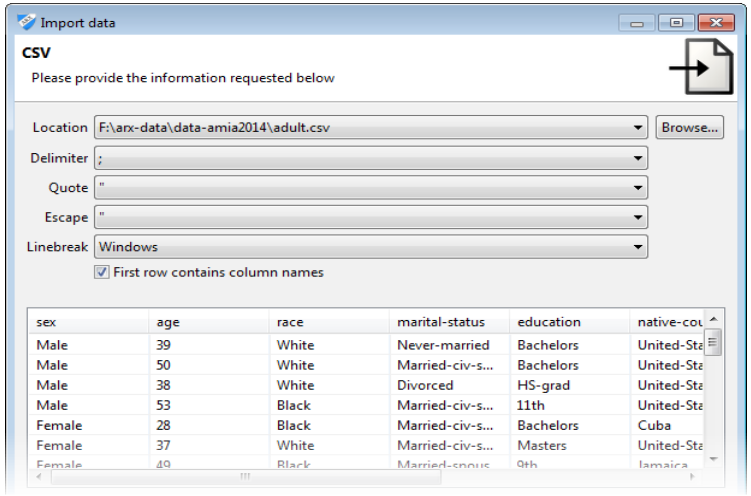


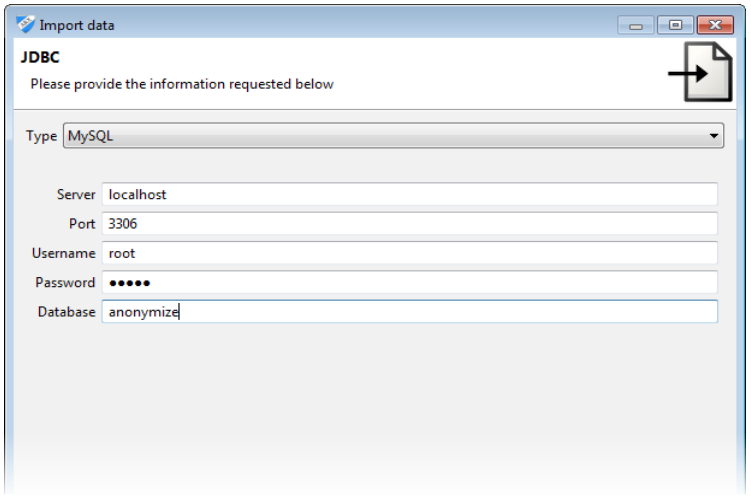
- Define transformation model
- Define privacy model
- Define coding model

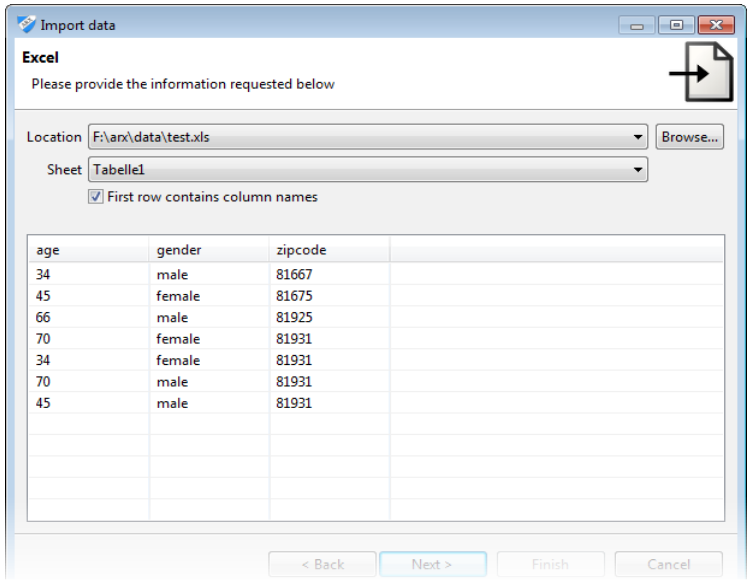
- Filter and analyze the solution space
- Organize transformations

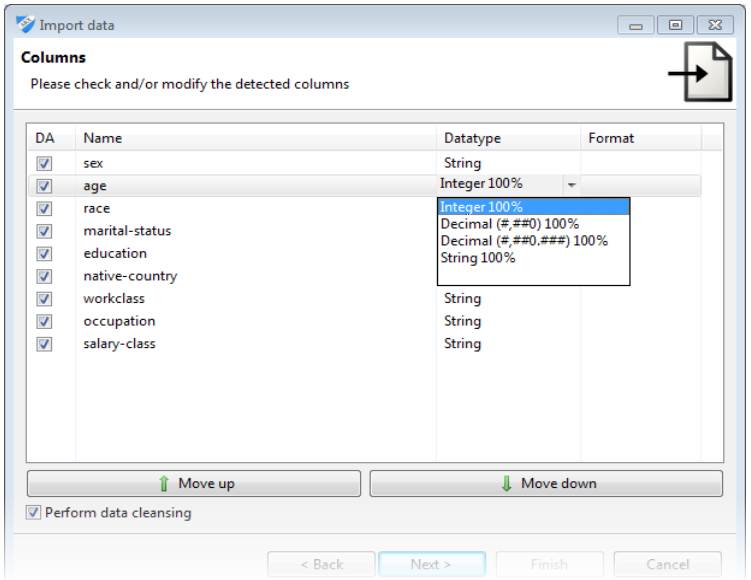
- Compare and analyze input and output
- Regarding risks and utility

ARX: Wizard for data import

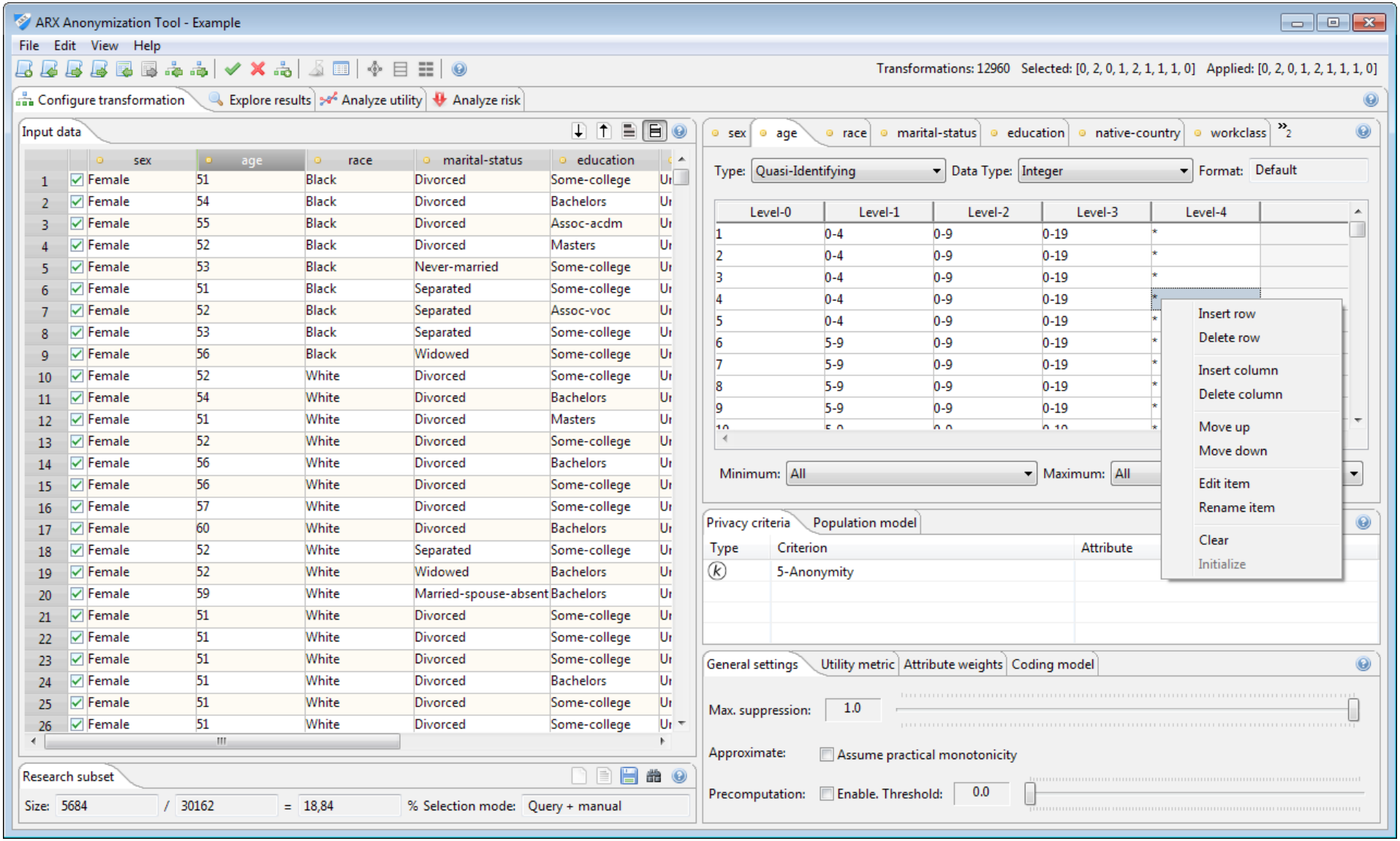




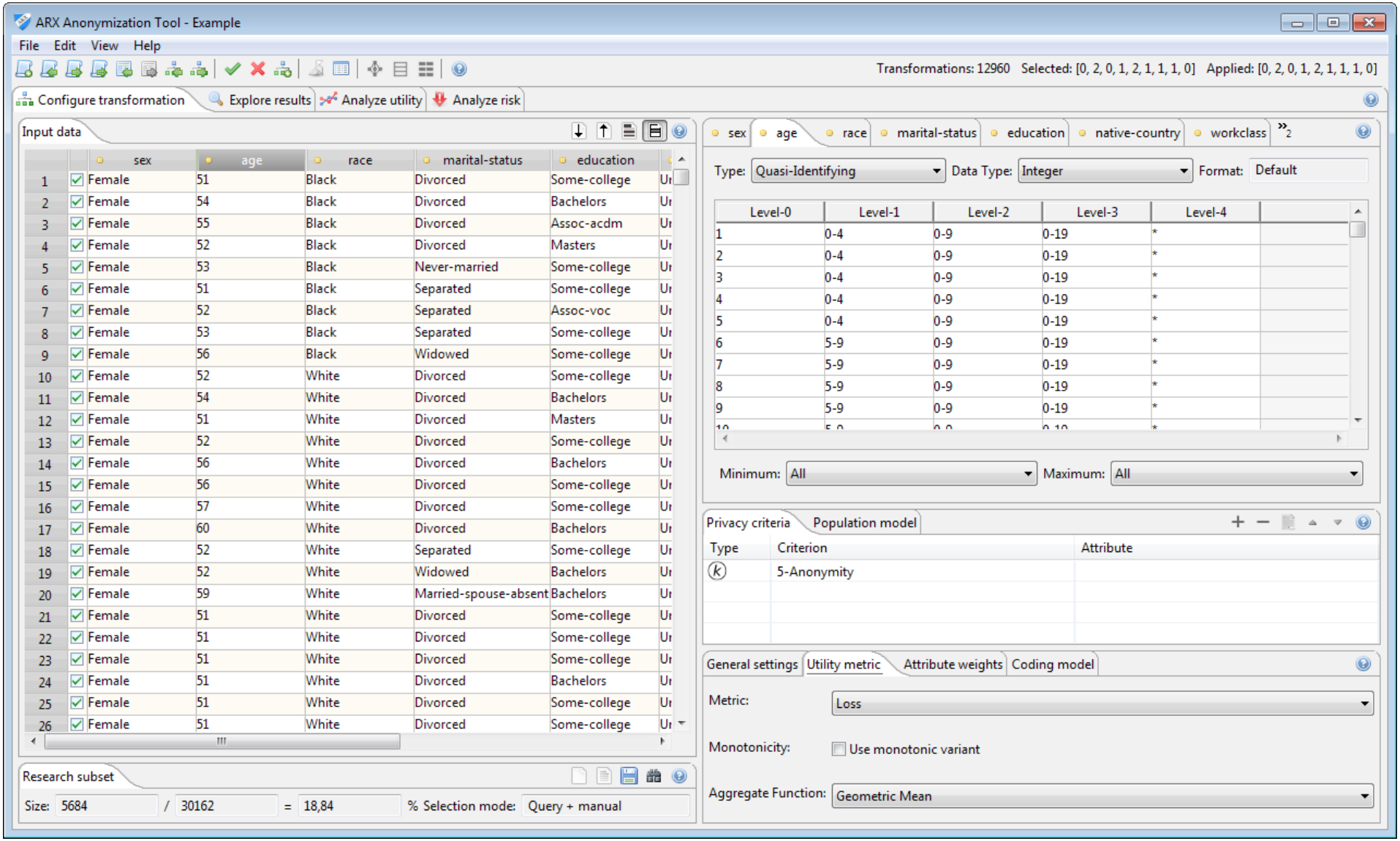




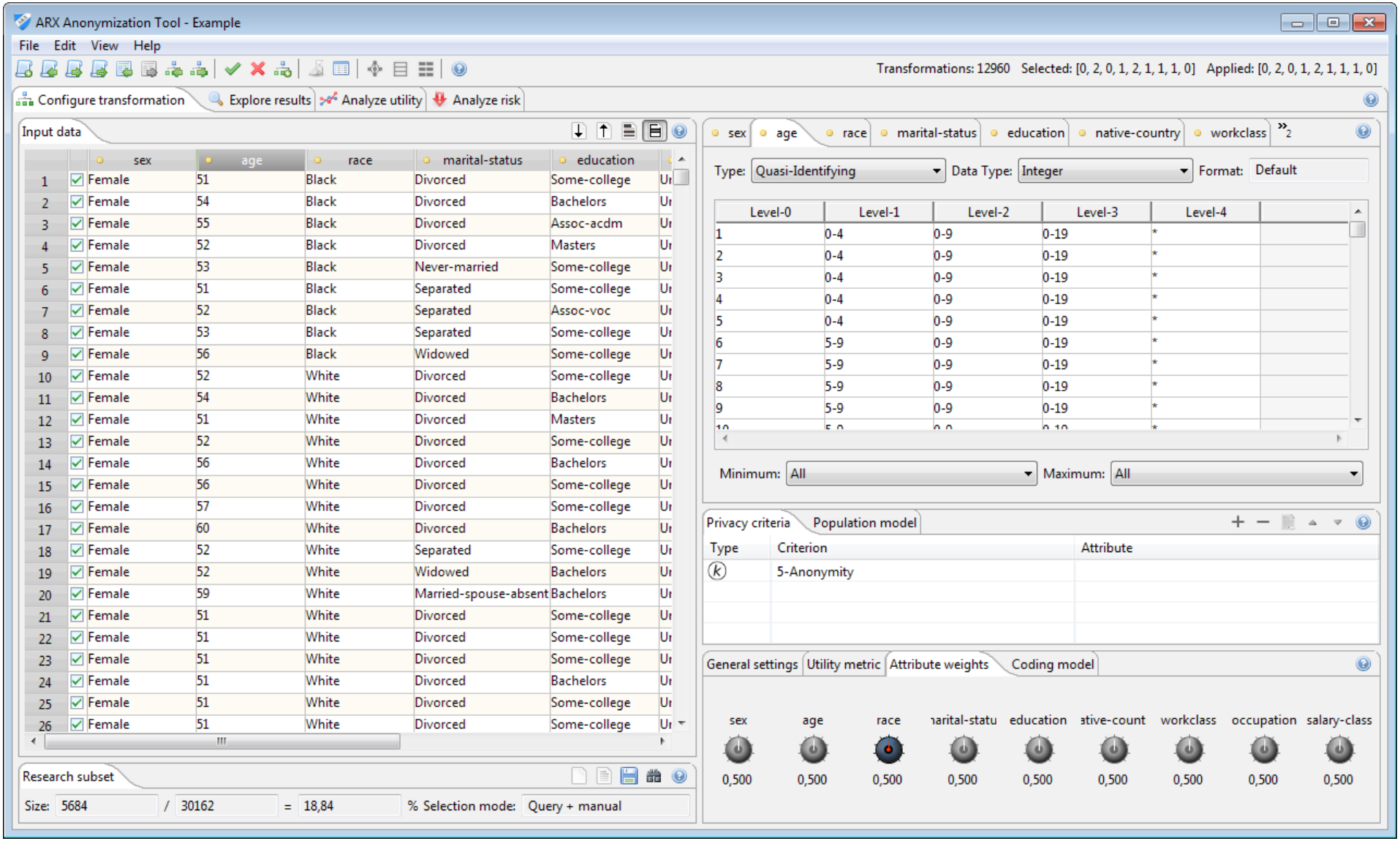
ARX: Configuration (1)



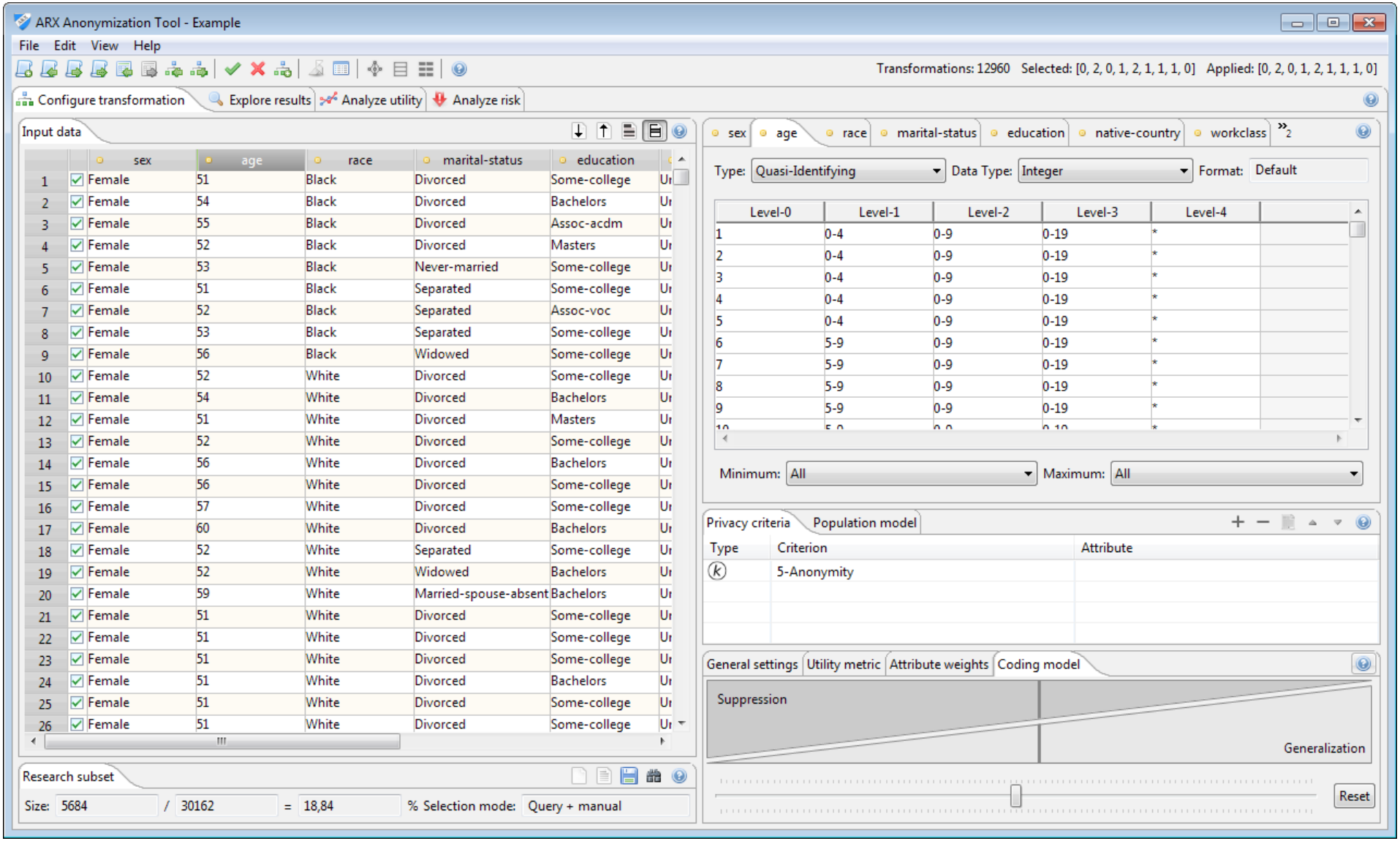
ARX: Configuration (2)



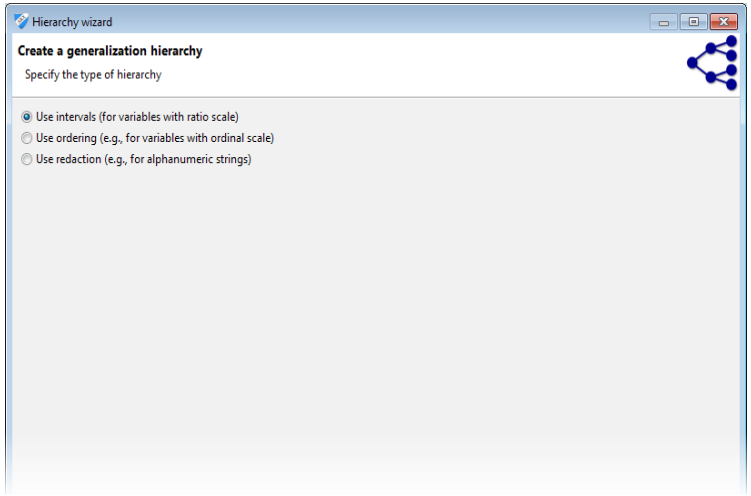
ARX: Configuration (3)

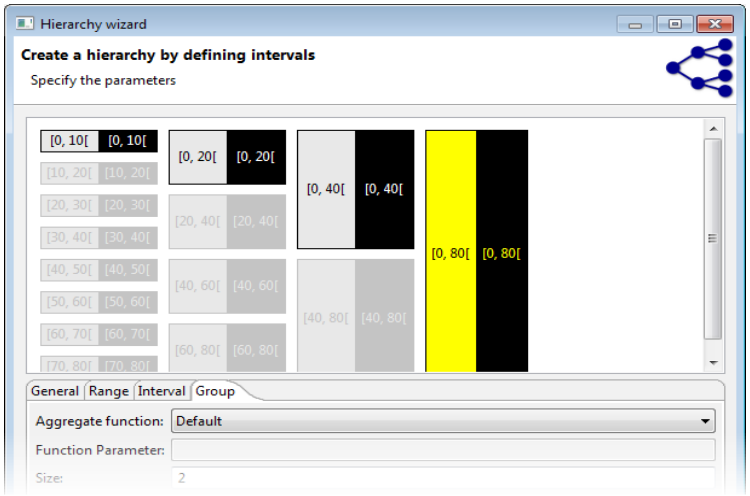


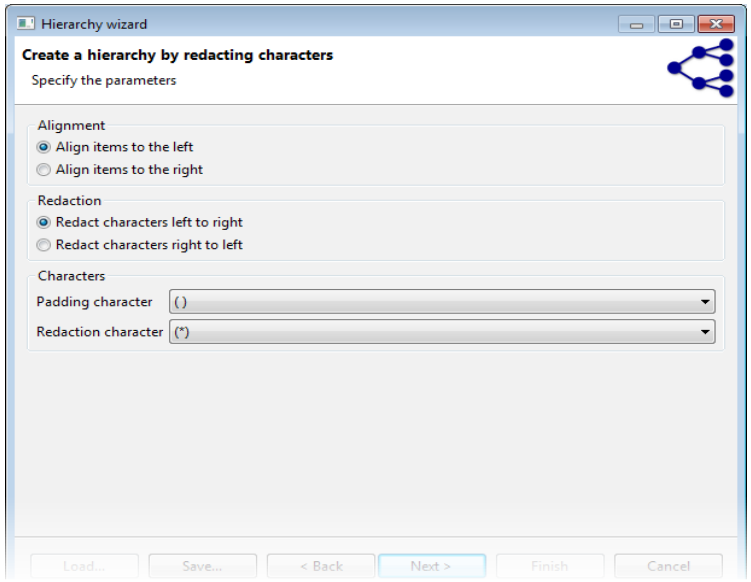
ARX: Configuration (4)

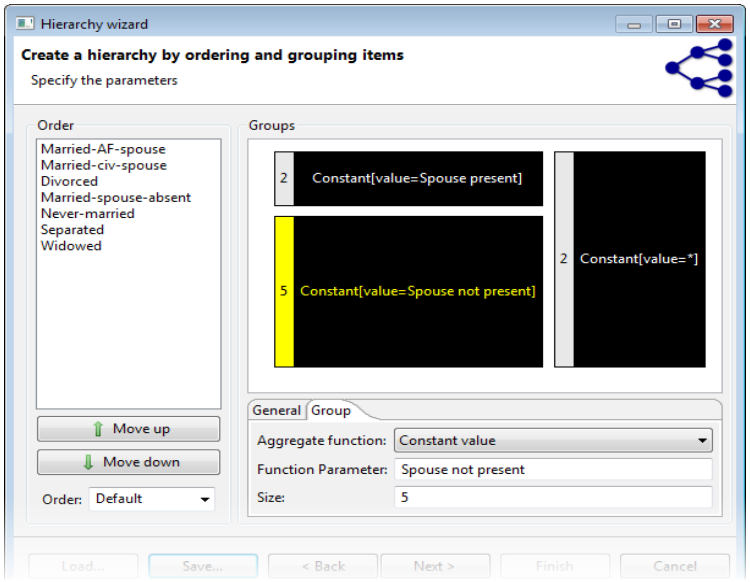


ARX: Wizard for transformation rules

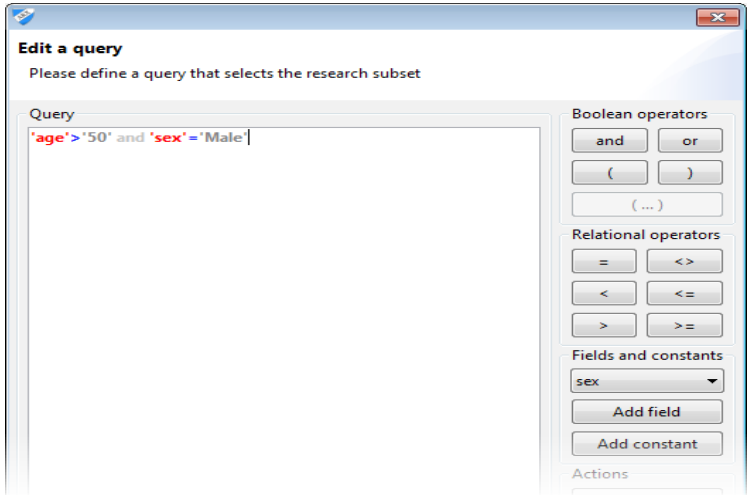


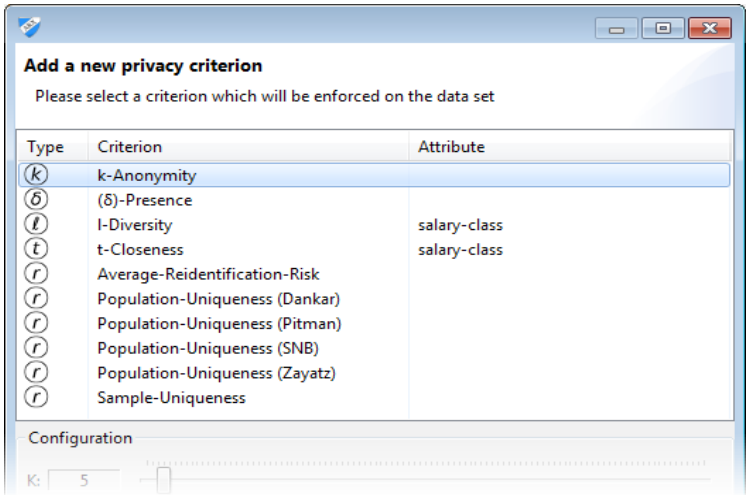




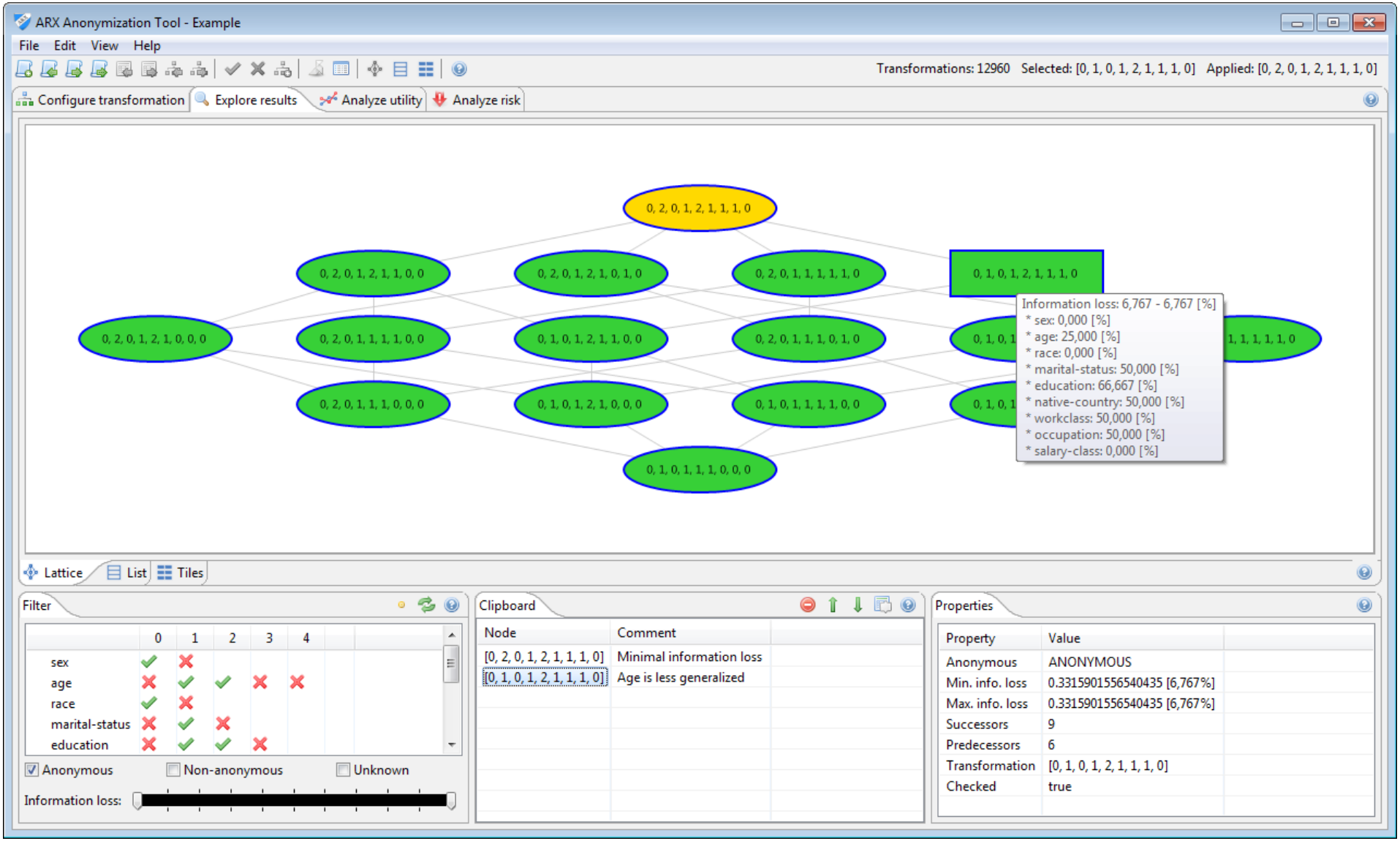


ARX: Further dialogs

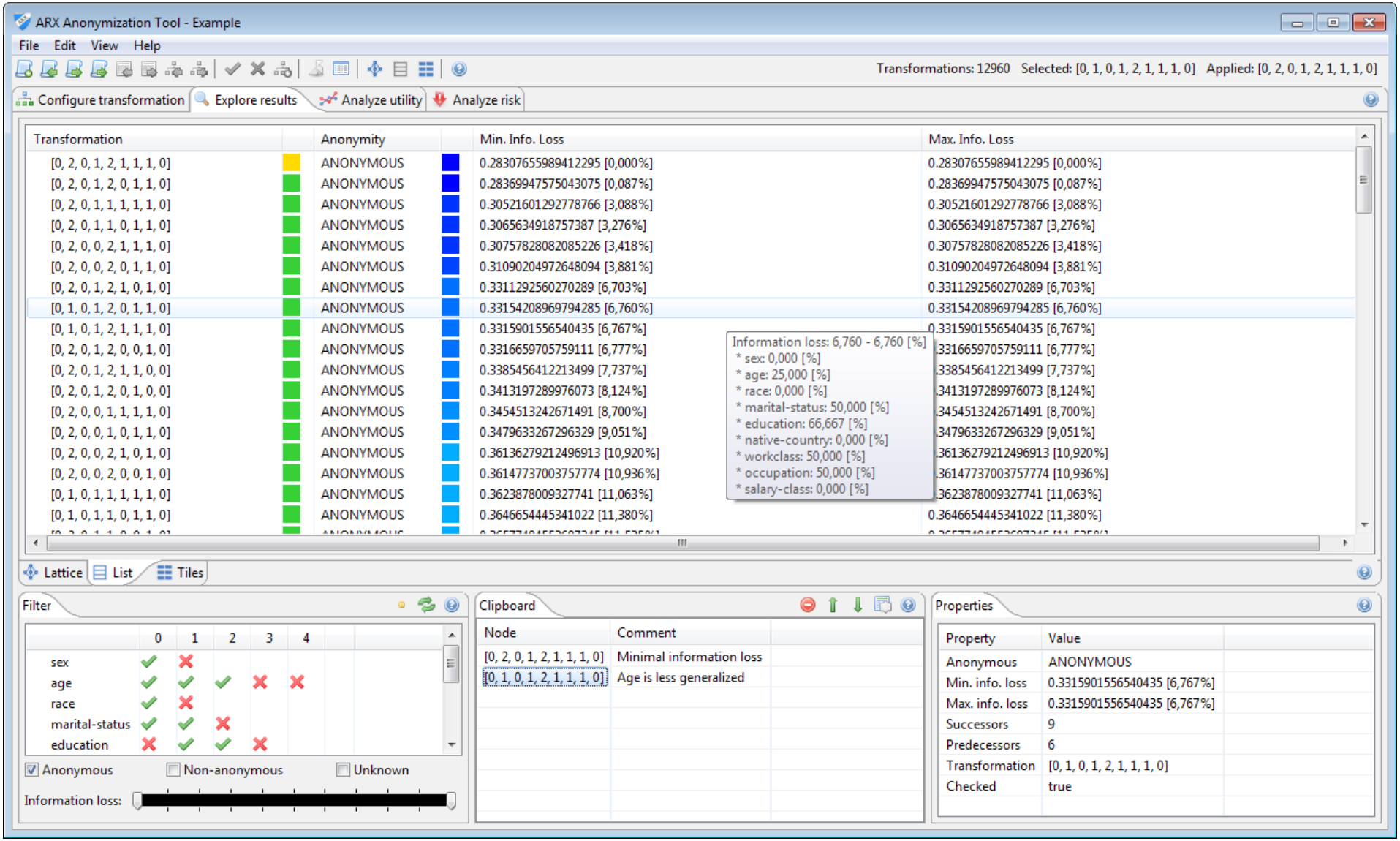




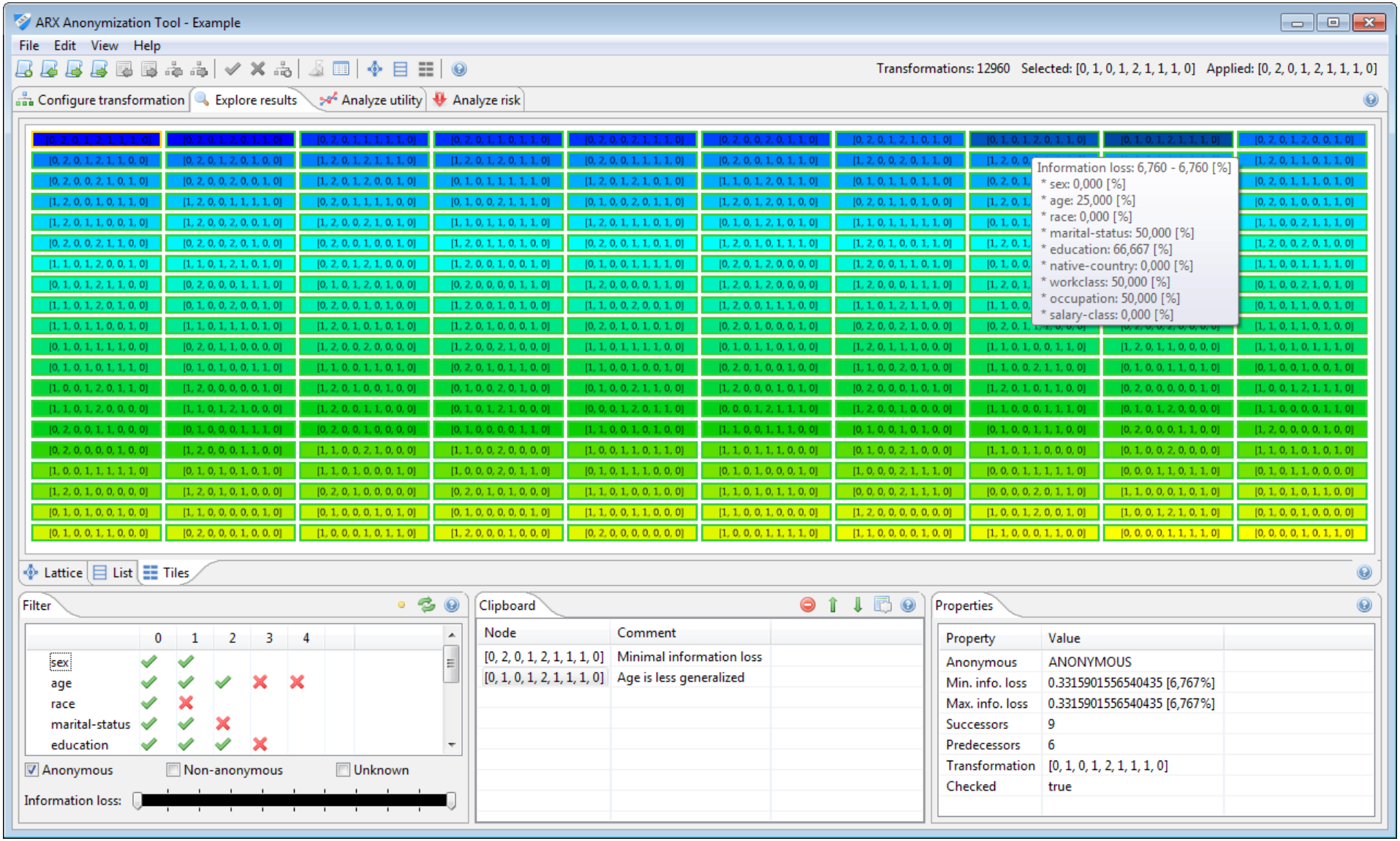
ARX: Exploration (1)



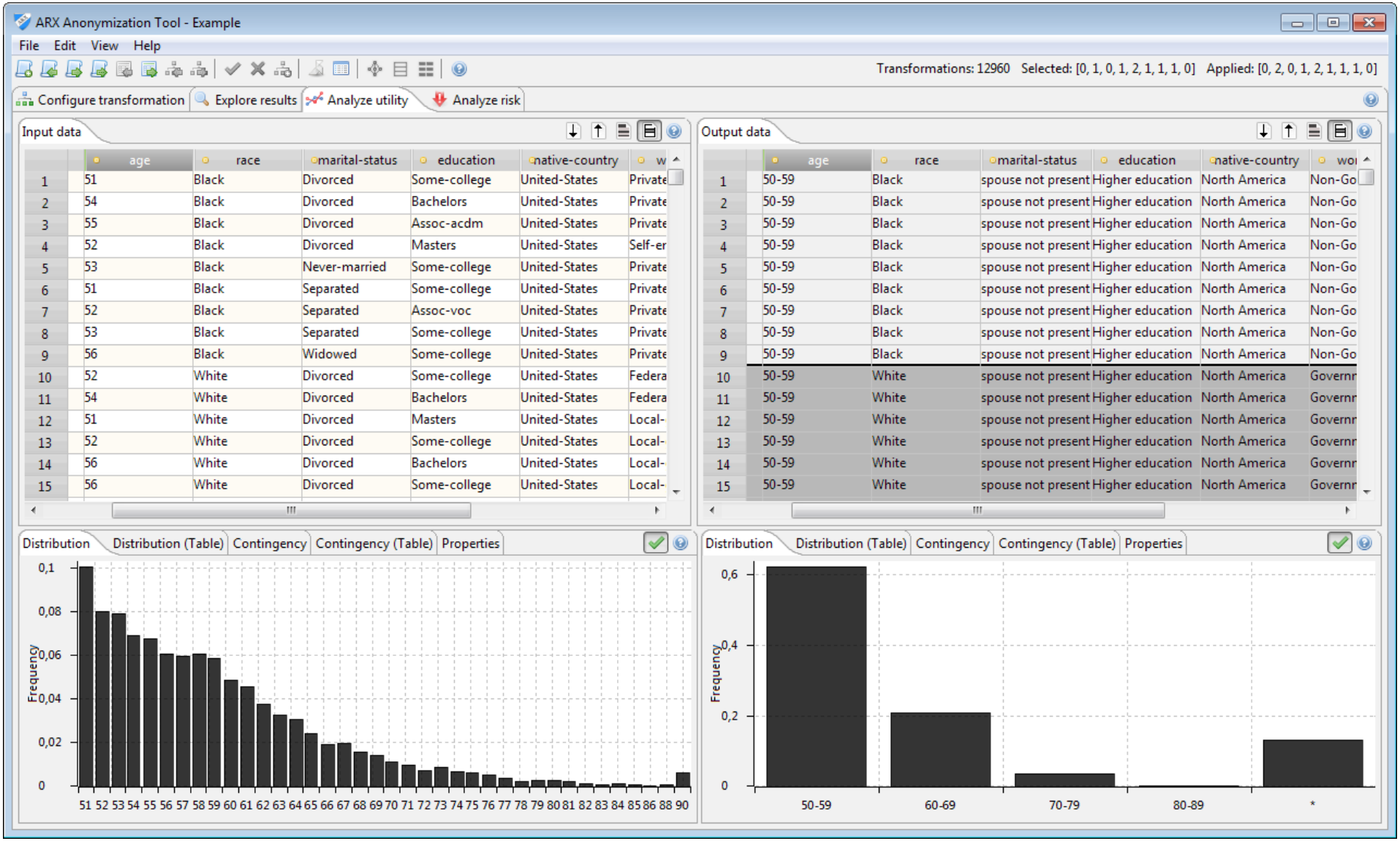
ARX: Exploration (2)



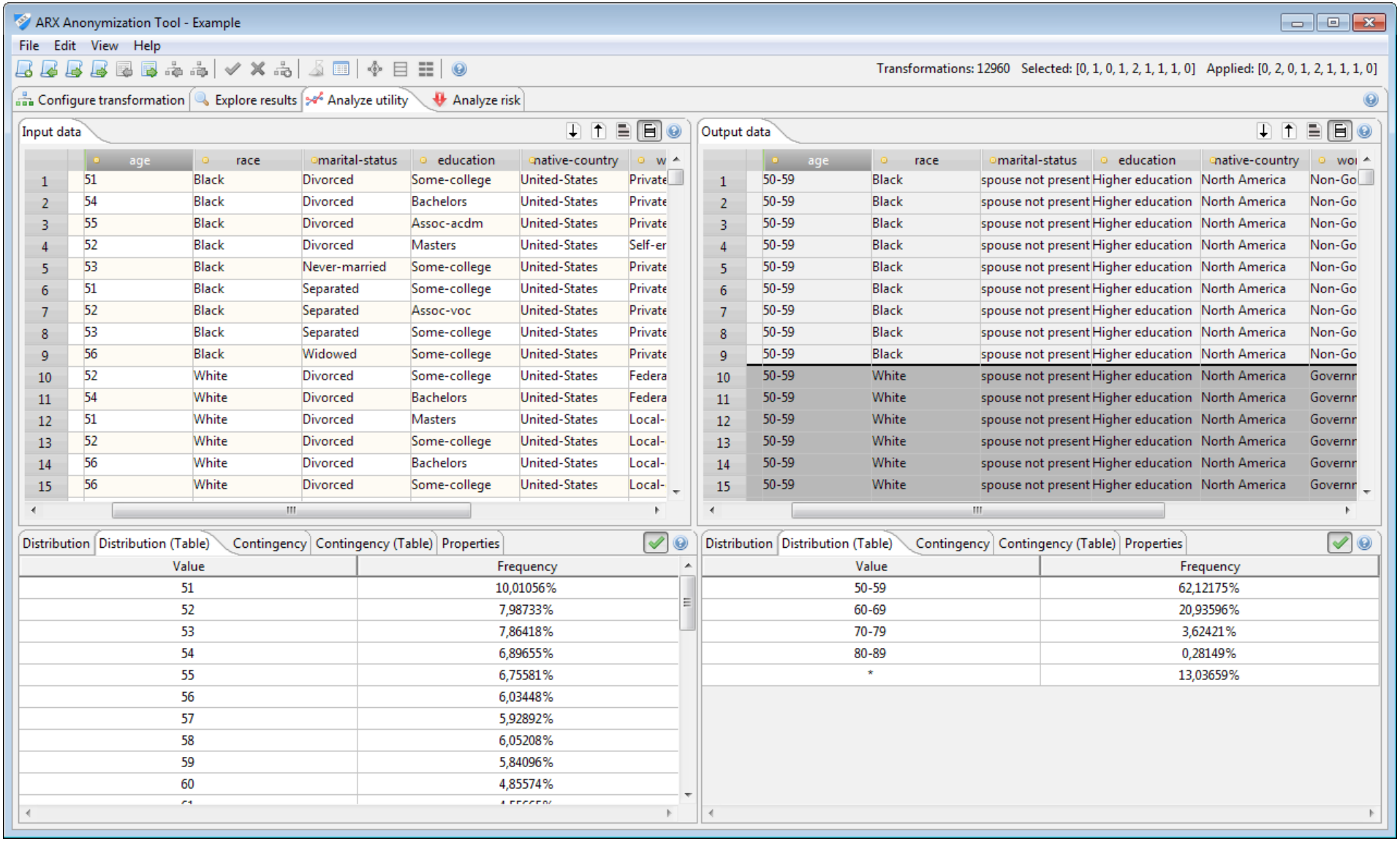
ARX: Exploration (3)



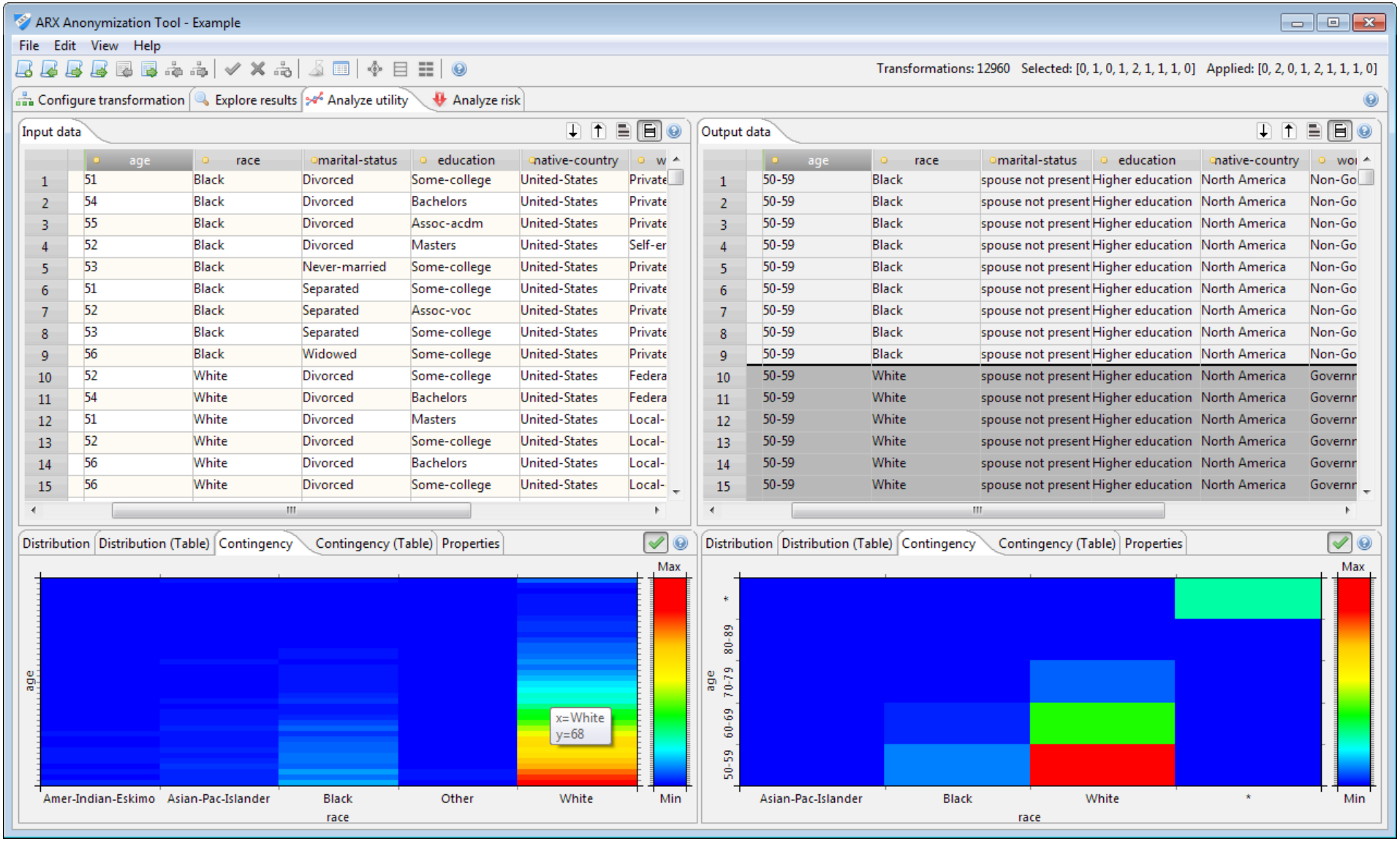
ARX: Utility analysis (1)



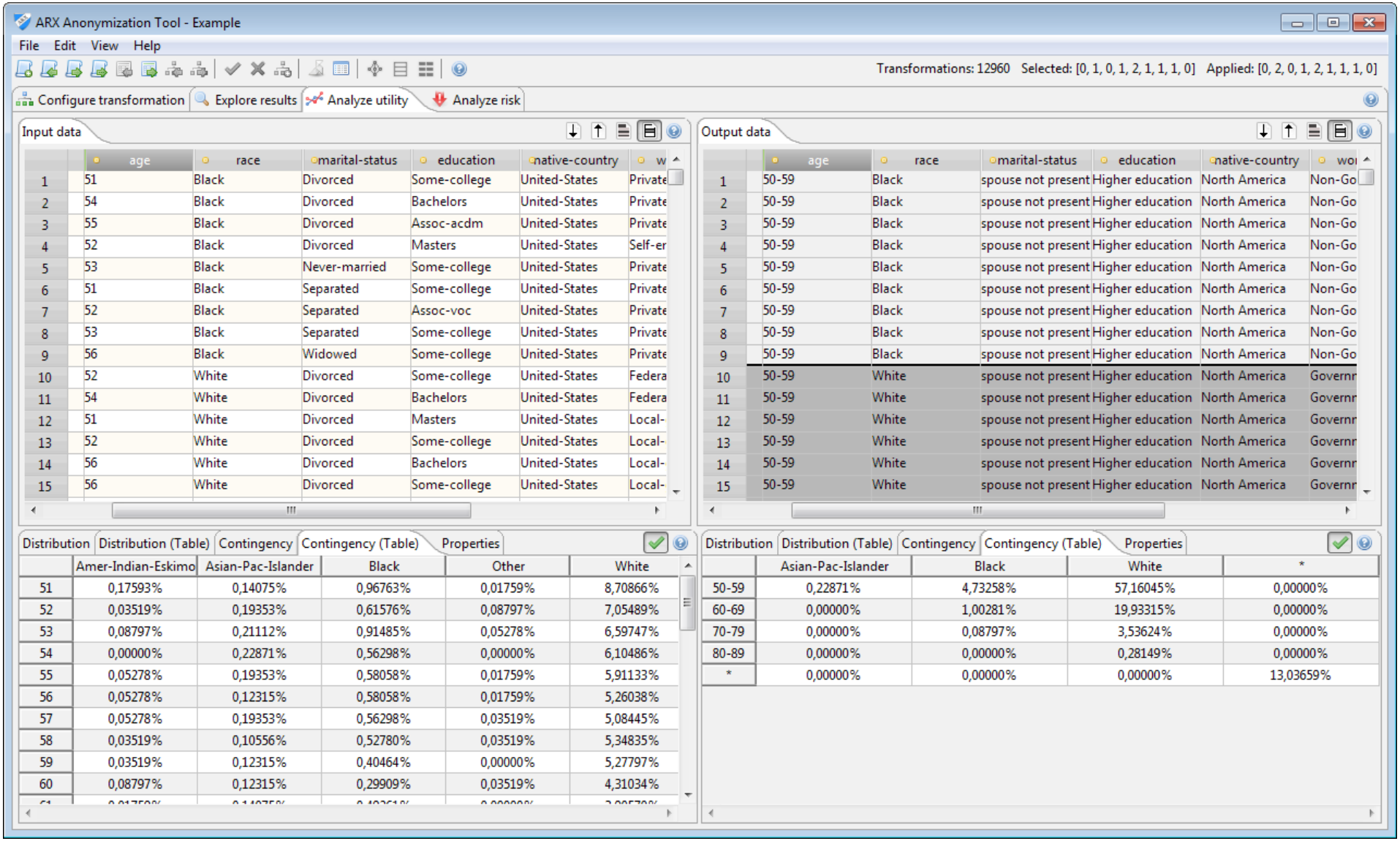
ARX: Utility analysis (2)



ARX: Utility analysis (3)



ARX: Utility analysis (4)



Input data

	age	race	marital-status	education	native-country	w
1	51	Black	Divorced	Some-college	United-States	Private
2	54	Black	Divorced	Bachelors	United-States	Private
3	55	Black	Divorced	Assoc-acdm	United-States	Private
4	52	Black	Divorced	Masters	United-States	Self-er
5	53	Black	Never-married	Some-college	United-States	Private
6	51	Black	Separated	Some-college	United-States	Private
7	52	Black	Separated	Assoc-voc	United-States	Private
8	53	Black	Separated	Some-college	United-States	Private
9	56	Black	Widowed	Some-college	United-States	Private
10	52	White	Divorced	Some-college	United-States	Federa
11	54	White	Divorced	Bachelors	United-States	Federa
12	51	White	Divorced	Masters	United-States	Local-
13	52	White	Divorced	Some-college	United-States	Local-
14	56	White	Divorced	Bachelors	United-States	Local-
15	56	White	Divorced	Some-college	United-States	Local-

Output data

	age	race	marital-status	education	native-country	wo
1	50-59	Black	spouse not present	Higher education	North America	Non-Go
2	50-59	Black	spouse not present	Higher education	North America	Non-Go
3	50-59	Black	spouse not present	Higher education	North America	Non-Go
4	50-59	Black	spouse not present	Higher education	North America	Non-Go
5	50-59	Black	spouse not present	Higher education	North America	Non-Go
6	50-59	Black	spouse not present	Higher education	North America	Non-Go
7	50-59	Black	spouse not present	Higher education	North America	Non-Go
8	50-59	Black	spouse not present	Higher education	North America	Non-Go
9	50-59	Black	spouse not present	Higher education	North America	Non-Go
10	50-59	White	spouse not present	Higher education	North America	Governr
11	50-59	White	spouse not present	Higher education	North America	Governr
12	50-59	White	spouse not present	Higher education	North America	Governr
13	50-59	White	spouse not present	Higher education	North America	Governr
14	50-59	White	spouse not present	Higher education	North America	Governr
15	50-59	White	spouse not present	Higher education	North America	Governr

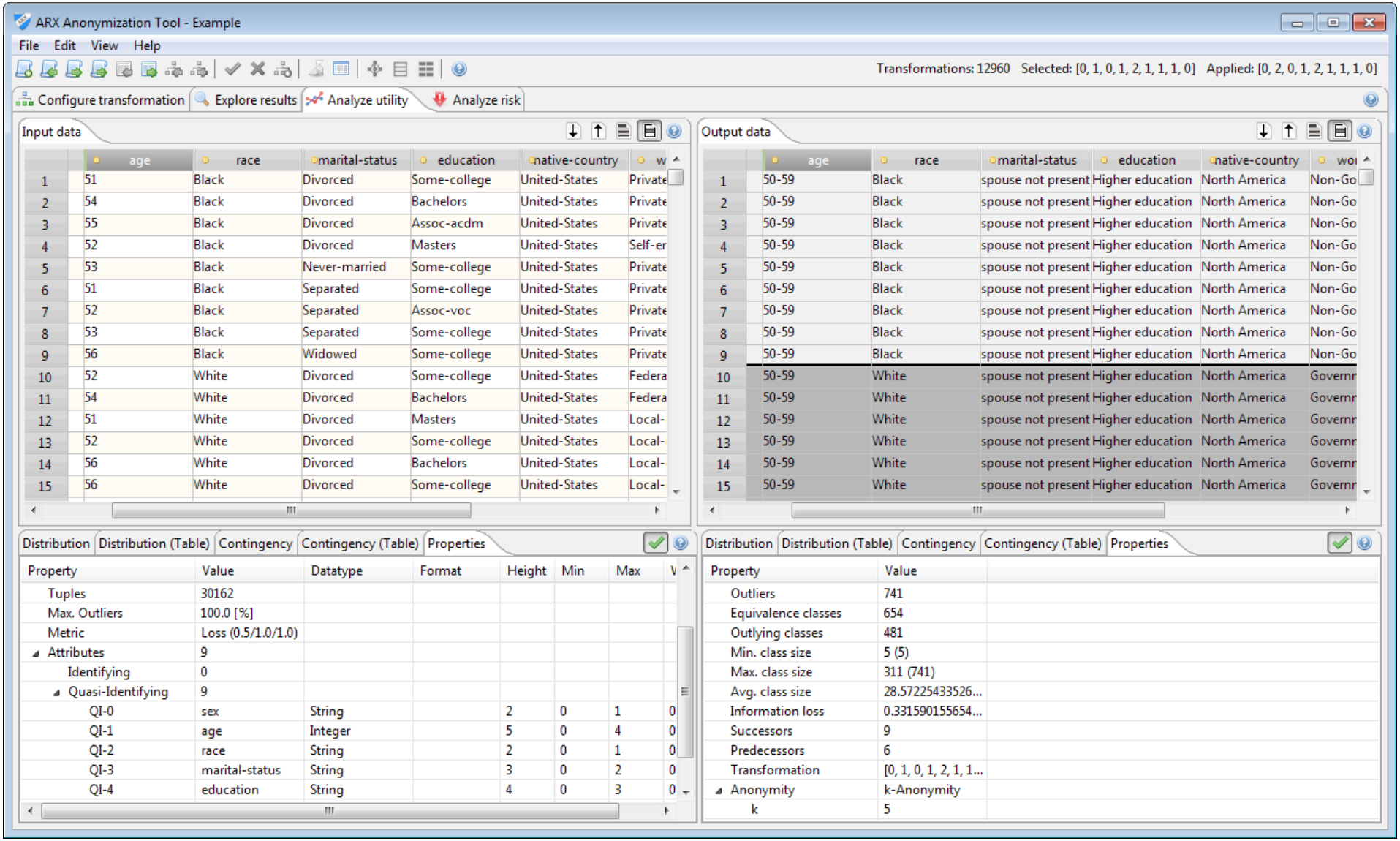
Distribution | Distribution (Table) | Contingency | Contingency (Table) | Properties

	Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White
51	0,17593%	0,14075%	0,96763%	0,01759%	8,70866%
52	0,03519%	0,19353%	0,61576%	0,08797%	7,05489%
53	0,08797%	0,21112%	0,91485%	0,05278%	6,59747%
54	0,00000%	0,22871%	0,56298%	0,00000%	6,10486%
55	0,05278%	0,19353%	0,58058%	0,01759%	5,91133%
56	0,05278%	0,12315%	0,58058%	0,01759%	5,26038%
57	0,05278%	0,19353%	0,56298%	0,03519%	5,08445%
58	0,03519%	0,10556%	0,52780%	0,03519%	5,34835%
59	0,03519%	0,12315%	0,40464%	0,00000%	5,27797%
60	0,08797%	0,12315%	0,29909%	0,03519%	4,31034%

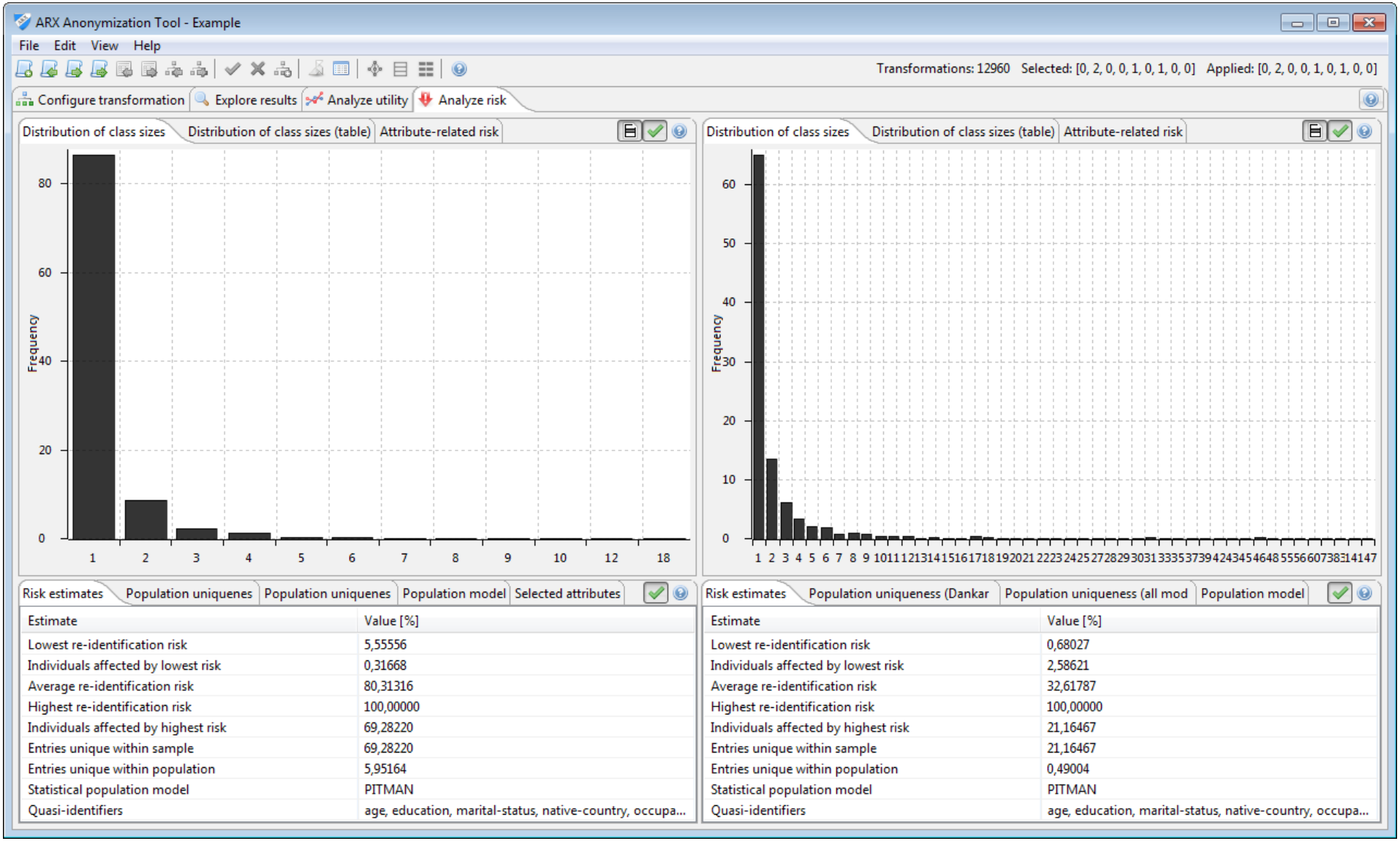
Distribution | Distribution (Table) | Contingency | Contingency (Table) | Properties

	Asian-Pac-Islander	Black	White	*
50-59	0,22871%	4,73258%	57,16045%	0,00000%
60-69	0,00000%	1,00281%	19,93315%	0,00000%
70-79	0,00000%	0,08797%	3,53624%	0,00000%
80-89	0,00000%	0,00000%	0,28149%	0,00000%
*	0,00000%	0,00000%	0,00000%	13,03659%

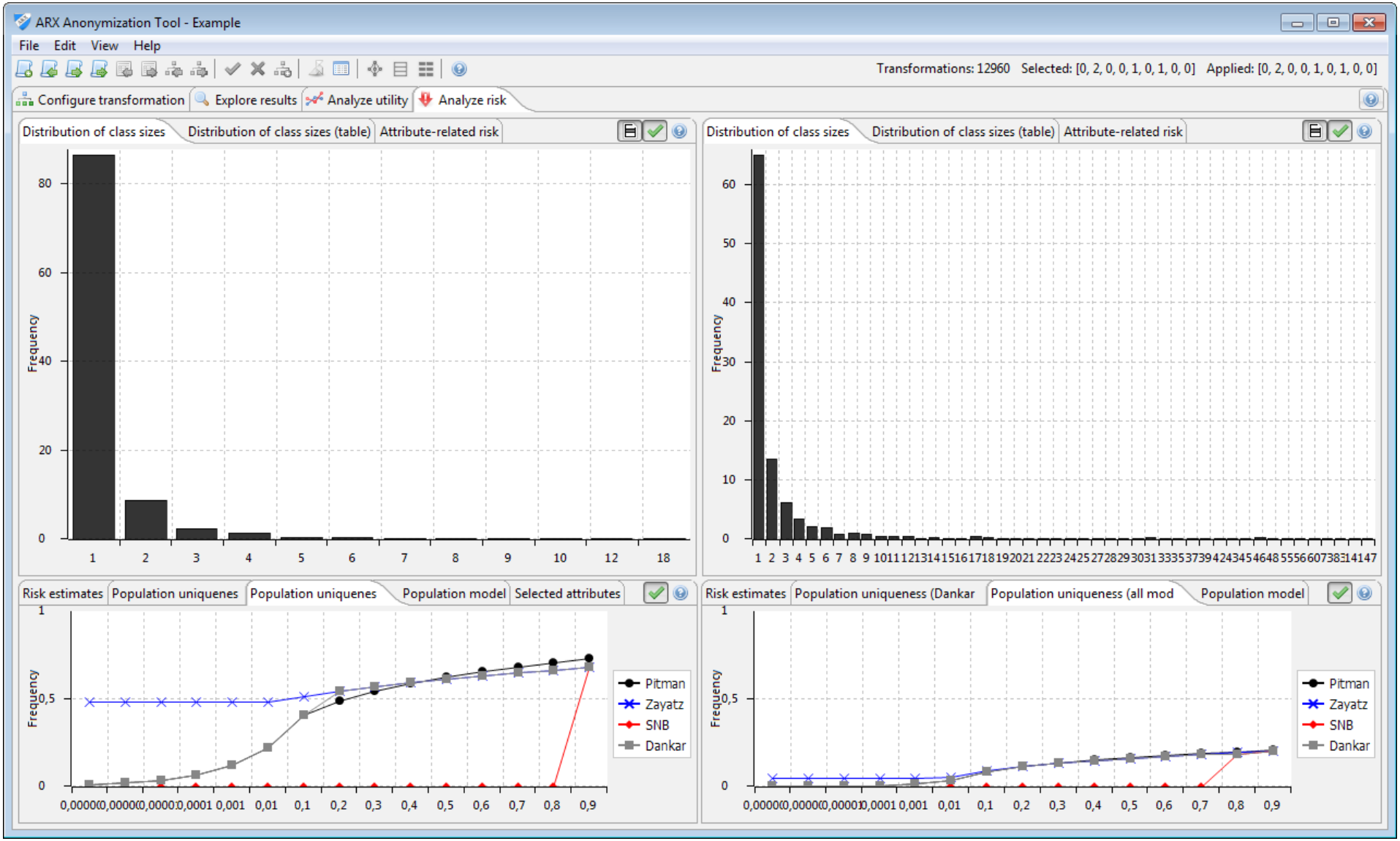
ARX: Utility analysis (5)



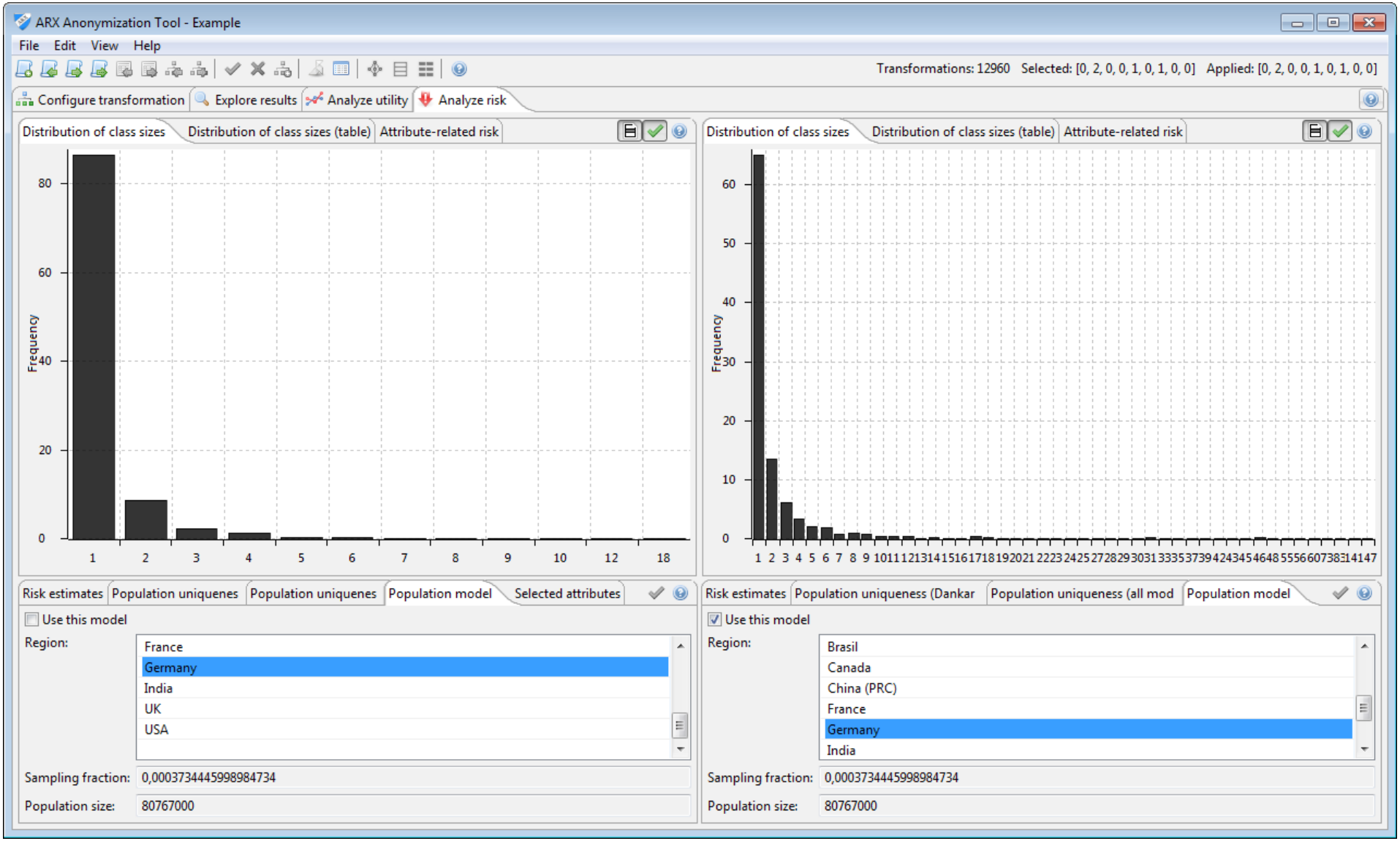
ARX: Risk analysis (1)



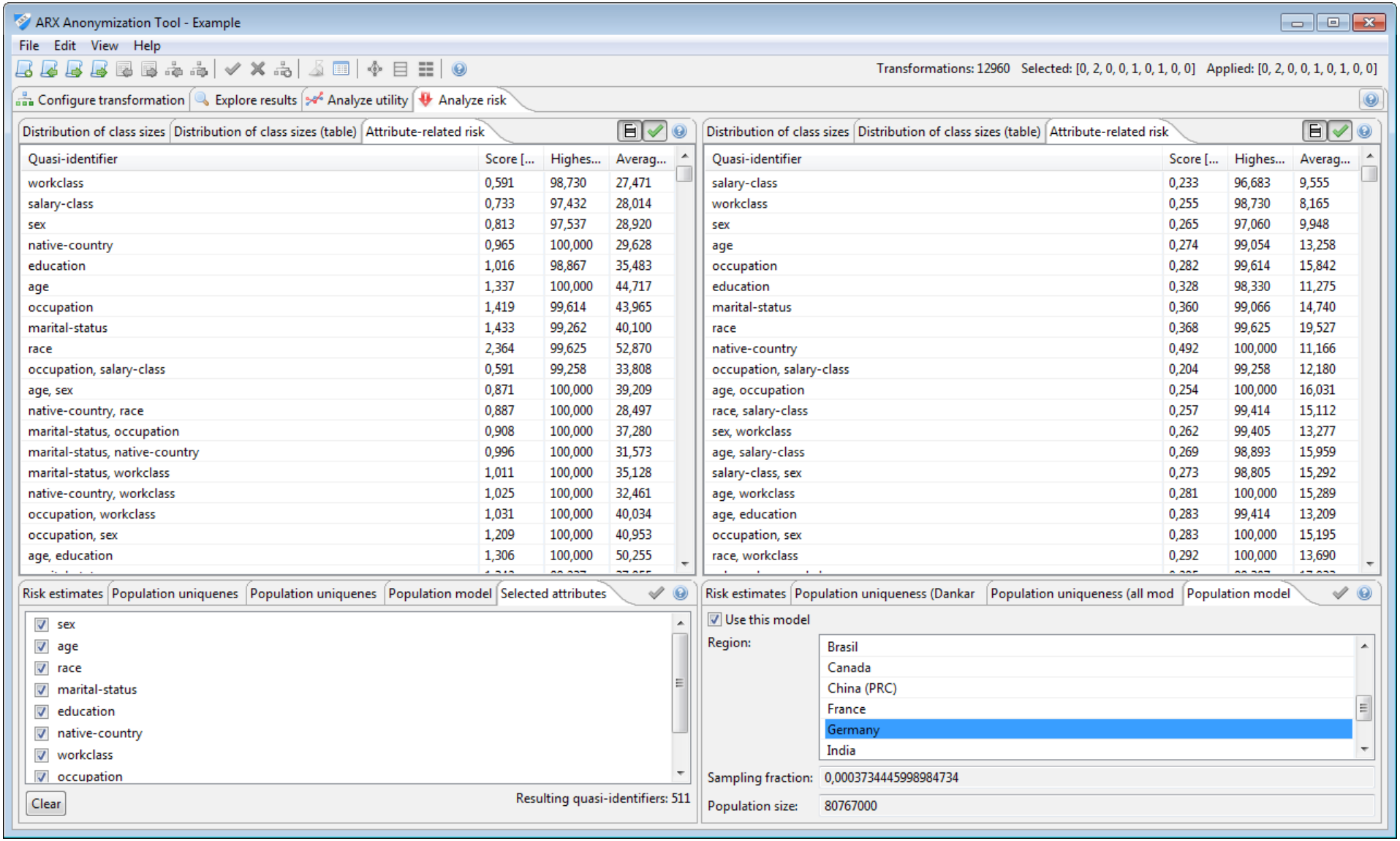
ARX: Risk analysis (2)



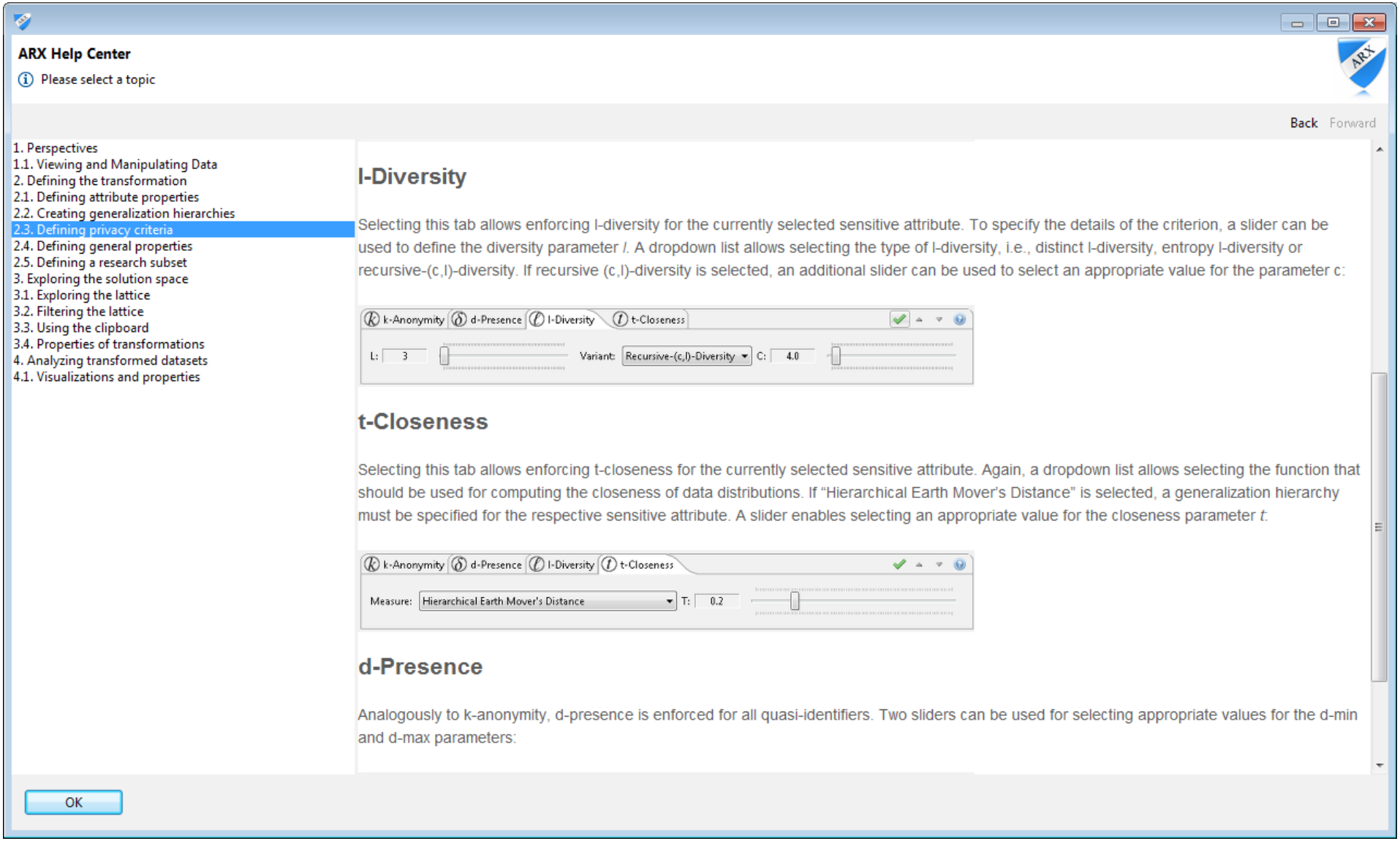
ARX: Risk analysis (3)



ARX: Risk analysis (4)



ARX: Context-sensitive help



ARX: Further developments

- **Current projects**
  - Non-interactive Differential Privacy
  - Support for high-dimensional data: heuristic algorithms
  - Further analyses and visualizations: utility and risks
  - Support for transactional attributes: (k, k^m)-anonymity
  - Integrated data masking methods
  - More flexible definition of quasi-identifiers
  - More risk models
- **Planned projects**
  - Auto-detection of HIPAA identifiers
  - Implement more flexible privacy criteria

Thank you for your attention

- **ARX is open source software**
- **Contribute:** feature requests, code reviews, criticism, enhancements, questions
- **Repository:** https://github.com/arx-deidentifier/arx

- **Further information & download:** http://arx.deidentifier.org

- **Get in touch**
  - Fabian Prasser (prasser@in.tum.de)
  - Florian Kohlmayer (florian.kohlmayer@tum.de)

- **Any questions?**