Advances in Social Science, Education and Humanities Research, volume 617
Proceedings of the International Conference on Social Science, Psychology and Legal Regulation (SPL 2021)

The Use of Sociological Methods in Criminological Research

Oleksandr Korystin [1, 2 * [0000-0001-9056-5475]], Nataliia Svyrydiuk [1 [0000-0001-9772-1119]], Alexander Vinogradov [3 [0000-0003-1250-3863]]

[1] *State Scientifically Research Institute of the Ministry of Internal Affairs, Kyiv, Ukraine*
[2] *National Security and Defense Council's of Ukraine Inter-Service Research Center into Combating Organized Crime Problems, Kyiv, Ukraine*
[3] *Taras Shevchenko National University of Kyiv, Ukraine*
** alex@korystin.pro*

ABSTRACT
The article focuses on the need to use sociological methods in criminological research. The current situation in Ukraine is characterized by inconsistency in state statistics on criminality in the country. And the only alternative can be an expert survey in this area and analysis of relevant data. This approach requires compliance with the scientific approach and achievement of representativeness of criminological conclusions. To this end, sociological science uses a variety of methods to achieve data reliability and conclusions validity. In order to test the hypothesis, the methods of logical error, cluster analysis were used; their efficiency for data purification and ensuring representativeness of criminological conclusions was analyzed.

Keywords: *criminological, sociological methods, expert survey, relevant data, conclusions validity, cluster analysis.*

1. INTRODUCTION

Modern law enforcement policies in developed security systems are mainly based on the principles of forming adequate knowledge about systemic criminal phenomena, criminal environment and development of appropriate strategies. At the same time, latent crime, in particular economic crime, as well as its organized manifestations, and in Ukraine the inadequacy of official criminal statistics on the level of criminogenic risks in society [1, 2], encourage the search for new forms of crime analysis. Examples of such approaches are Europol's SOCTA methodology [3] and risk-oriented law enforcement strategies [4, 5], based on expert surveys and the use of appropriate sociological and statistical methods [6, 7] of analysis.

Data collection using online surveys is becoming an increasingly common means of obtaining information from respondents, given the ability to quickly learn the opinions of a large number of people who are difficult to reach on a large number of questions.

However, the quality of data collected from such surveys can suffer significantly due to insufficient expertise, reduced motivation to provide thoughtful answers, fatigue and inattention due to excessive number and complexity of questions and distractions, installation behavior, quality of tools and technical aspects of the survey.

2. RESEARCH ANALYSIS

To control the installation styles and attentiveness when filling out questionnaires, researchers usually include in the questionnaire specially created control questions [8]:

- to measure the social desirability of answers;

- identical in content questions with forward and reverse key;

- questions to check the competencies of respondents;

- direct instructions (for example: "Choose answer #2 to this question").

If such tools were not planned before the survey, it is necessary to identify inattentive, incompetent or unmotivated respondents in another way, based on the statistical properties of their answers.

The following signs of low data quality in a particular respondent may be [9]:

- a significant number of passes;

- too short response time to questions;

- low variability of answers to a number of questions;

- presence of non-random patterns of answers to a number of questions that are not determined by their content;

- significant difference between the answers of the respondent and the profile of the answers of a significant number of bona fide and competent participants.

3. THE PURPOSE OF THE ARTICLE

Predicated on the unique empirical base formed by the survey of the subjects of fiscal security of Ukraine [5], to suggest data cleansing approaches of general expert set, forming a reliable expert sample [8] and ensuring sufficient representativeness of analytical conclusions. Analyze the spread risks of targeted activities of organized crime in the field of fiscal security of Ukraine.

4. THE MAIN MATERIAL

Already initial analytical conclusions, conducting a strategic analysis in the field of fiscal security of Ukraine based on a risk-oriented approach, focused on the extreme importance of such innovation in Ukraine and formed a system of knowledge to justify priority areas for further development of this segment in Ukrainian society.

Using the risk management methodology, threats in the field of fiscal security were identified and assessed (243), analysis and risks assessment of their spread were carried out, risk limits were determined, abilities (119) and external opportunities (52) of risk minimization were assessed, and a forecast model was designed to manage these risks [5].

The most significant threats, with a risk level exceeding 60 % of the maximum size and requiring urgent measures to reduce the risk of their spread are VAT fraud and lobbying by government officials of certain groups and companies to promote favorable tax (customs) preferences [5].

At the same time, some threats, although characterized by a slightly lower level of risk, still require control of top management of key fiscal security actors: targeted activities of organized crime in the fiscal sphere (57%), pressure on business (57%), "schemes" of minimization tax payments (56.6%), smuggling (56%), shadowing of certain types of business (55%), avoidance of criminal liability for tax offenses (55%) and manifestations of corruption in the fiscal sphere (55%) [5].

In this case, it is important to note that the conclusions obtained require a sufficient level of objectivity, reliability and representativeness of the data. That is why several ways to clear the collected data set were used.

Our hypothesis is based on the assumption of insufficient reliability of the data obtained in the survey of more than 7,000 respondents (experts in the field of fiscal security). This assumption was motivated by the fact that respondents had to fill in the questionnaire online in two stages in an online mode without external supervision and evaluate the "likelihood" and "consequences" of more than 400 indicators [5], which characterized threats, abilities (vulnerabilities) and external opportunities. Of course, this task is difficult and not all respondents could pay enough attention to filling out the questionnaire. The statistical substantiation of the sampling restriction procedure is based on the fact that due to the large volume of the questionnaire, experts could make mistakes in the answers, as the complexity of the questions and the limited time of their comprehension leads to instability of attention [10].

In order to test the hypothesis and ensure the probable purification of the data, first of all, the method of logical error was used. In particular, separate indicators in different sections of the questionnaire were used to check the competencies of the respondents. For example, according to the "*Shadow Economy Level*" indicator, a reasonably reasonable answer when estimating the likelihood of spreading this threat on a "low-medium-high" scale is possible "medium" or "high" probability, but the answer "low likelihood" will not be considered competent level of the shadow economy in Ukraine.

Using this indicator as a filter, the first sample was made from the data set of the general population, which significantly improved the quality of data for further analysis. This can be seen in the following example: the level of risk of spreading the threat according to the TA2 indicator (*targeted activity of organized criminal groups in relation to VAT*), which is one of the biggest threats, has the following distributions in the group of those selected by the no-error filter who did not pass this filter (Table 1):

Table 1. TA2 * filter_$ Crosstabulation

		filter_$		Total
		Not Selected	Selected	
TA2	low	16.7%	2.2%	11.5%
	medium	43.2%	17.8%	34.1%
	high	40.1%	80.0%	54.4%
Total		100.0%	100.0%	100.0%

As can be seen, the difference in distributions is significant: 16.7% of unreliable experts indicated a low likelihood of a threat, while reliable ones chose this option in only 2.2% of cases. The "high likelihood" option was chosen by them in 80% of cases. This difference is not only statistically significant (criterion $\chi^2 = 38.378$, $p < 0.001$), but also the magnitude of the effect is very significant (V Cramer = 0.39). Similar trends are observed in other important questions of the questionnaire.

To improve the quality of the selected data, which, in our opinion, is a reasonable assumption, we used a technique based on identifying patterns (profiles) of respondents' answers to questions followed by determining the quality of information in the profile based on variability of answers, missed data and meaningfulness answers (this is determined separately). To identify groups of respondents with different styles of responding to a number of questions, we used the following algorithm, which is based on cluster analysis [11, 12, 13] by the method of K means.

The quality of the information provided by the respondent is determined separately for the selected block of questions, and not for the questionnaire as a whole. 35 indicators (TA2-TA36) were analyzed, which characterize the level of risk of spreading threats related to targeted activities of organized crime in the field of fiscal security. This approach makes it possible to preserve the maximum of expert information, as it is assumed that the attentiveness, competence and motivation of the expert may vary over time or depend on the content of questions in the block. It should also be noted that the level of risk is determined on the basis of an integrated assessment of the likelihood and consequences of identified threats.

Statistical analysis of the data was performed using the IBM SPSS Statistics version 25, but the proposed algorithm is easy to implement using the R language [14].

The next step in the algorithm is to fill in the missing answers. It should be noted that a significant amount of missed data is a significant problem in any statistical analysis, so the gaps (in this case – code 99 is) were replaced by the authors with a value close to the modal response (e.g. 10), but slightly different from it (e.g. 9) to facilitate the interpretation of clusters (Figure 1).

```
FREQUENCIES VARIABLES=TA2 to TA36
  /FORMAT=NOTABLE
  /STATISTICS=MEAN MEDIAN
  /ORDER=ANALYSIS.
RECODE TA2 (MISSING = 25).
RECODE TA3 (MISSING = 25).
RECODE TA4 (MISSING = 25).
RECODE TA5 (MISSING = 15).
RECODE TA6 (MISSING = 25).
RECODE TA7 (MISSING = 10).
RECODE TA8 (MISSING = 15).
RECODE TA9 (MISSING = 15).
RECODE TA10 (MISSING = 15).
RECODE TA11 (MISSING = 15).
RECODE TA12 (MISSING = 15).
RECODE TA13 (MISSING = 30).
RECODE TA14 (MISSING = 30).
RECODE TA15 (MISSING = 30).
RECODE TA16 (MISSING = 30).
RECODE TA17 (MISSING = 30).
RECODE TA18 (MISSING = 30).
RECODE TA19 (MISSING = 30).
RECODE TA20 (MISSING = 25).
RECODE TA21 (MISSING = 30).
RECODE TA22 (MISSING = 30).
RECODE TA23 (MISSING = 30).
RECODE TA24 (MISSING = 30).
RECODE TA25 (MISSING = 30).
RECODE TA26 (MISSING = 30).
RECODE TA27 (MISSING = 30).
RECODE TA28 (MISSING = 15).
RECODE TA29 (MISSING = 30).
RECODE TA30 (MISSING = 30).
RECODE TA31 (MISSING = 15).
RECODE TA32 (MISSING = 15).
RECODE TA33 (MISSING = 30).
RECODE TA34 (MISSING = 15).
RECODE TA35 (MISSING = 15).
RECODE TA36 (MISSING = 15).
execute.
```

Figure 1 IBM SPSS Statistics syntax for determining the median and replacing missing data

This allowed the authors to identify the lack of answers in the pattern.

The next step is a cluster analysis of the data set on the selected block of questions by one of the selected algorithms (Two Step Cluster Analysis, K-Means

Clustering). Experiments have shown that the most convenient to use was the method of K-means.

It is important to note that the results of clustering are always somewhat unstable and depend on which variables the array was sorted by. Therefore, the clustering procedure should be repeated several times, pre-sorting the array by different variables and checking the stability of the solution.

The choice of the number of clusters into which the whole set is divided are important points in the clustering. At the initial stage, given the total sample of experts, we identified 10 clusters (Figure 2).

```
QUICK CLUSTER TA2 TO TA36
 /MISSING = LISTWISE
 /CRITERIA = CLUSTER(10) MXITER(999) CONVERGE(0)
 /METHOD = KMEANS(NOUPDATE)
 /PRINT INITIAL.
```

Figure 2 Syntax of cluster analysis by the method of K means

At the same time, some of the clusters turned out to be sparsely filled and somewhat similar, which indicates the need to gradually reduce their number to the optimal one. In our data sample, the choice of three clusters is optimal (Table 2).

Table 2. Descriptive Statistics

	N	Min	Max	Mean	Std. Deviation
K_1	35	70.15	89.70	81.3793	4.75563
K_2	35	26.12	70.58	46.9253	13.27791
K_3	35	18.75	35.42	26.2580	4.93323
Valid N (listwise)	35				

As a result of clustering, the variables of the respondent to the cluster and distance to its center are preserved in the main array.

After studying the variability and meaningfulness of clusters, we identified a cluster K2, the variability of the level of risk is estimated from min - 26.12% to max - 70.58%.

In the future, we formed a new variable in the data array "QCL_1" and based on the use of the filter on the cluster K2 data sampling is limited in order to improve it (Table 3).

Table 3. QCL_1 * filter_$ Crosstabulation

% within filter_$

		filter_$		Total
		Not Selected	Selected	
QCL_1	1	17.0%		11.2%
	2		100.0%	34.3%
	3	83.0%		54.5%
Total		100.0%	100.0%	100.0%

Thus, only 34.3% of the data were included in the sample limited by the filter-defined cluster K2. This limitation is not only statistically significant (criterion χ^2 = 268.000, p <0.000), but also the magnitude of the effect is very significant (V Cramer = 1.0).

These conclusions confirm our hypothesis about the poor quality of the general survey in the field of fiscal security of Ukraine. Therefore, the use of such data requires a mandatory procedure for their purification.

Further application of the identified sampling restriction filters provides an opportunity to finally assess the risks on 35 indicators of threats to the spread of organized crime in the field of fiscal security of Ukraine (Table 4).

The final sample of data based on risk assessment identifies four groups of threats to the spread of targeted organized crime in the field of fiscal security: the most significant of them (TA2, TA16, TA17, TA18, TA21, TA22); significant (TA3, TA4, TA6, TA13, TA14, TA15, TA19, TA25, TA27, TA30).

Table 4. Risk assessment of targeted activities of organized crime in the field of fiscal security of Ukraine

VARIABLE	**RISK ASSESSMENT, %**		
	WITHOUT FITTERS	**FITTER 1**	**FITTER 1,2**
TA2	48.5350	55.6219	**68.9082**

TA3	35.5683	46.1607	54.2667
TA4	40.5310	50.7016	59.9570
TA5	32.0987	42.4195	46.8576
TA6	40.9139	50.1681	56.3230
TA7	31.1044	39.4476	38.2219
TA8	28.5332	37.2112	35.2101
TA9	30.2258	36.2145	32.6282
TA10	29.0705	34.3430	29.2018
TA11	30.9699	37.7759	35.7203
TA12	29.6228	37.9104	34.6382
TA13	39.1844	46.6739	54.0850
TA14	36.9779	44.4832	51.3303
TA15	46.8267	48.4421	55.6295
TA16	49.6172	51.2993	**63.3274**
TA17	48.0425	52.5986	**66.2784**
TA18	41.4350	49.4843	**62.7223**
TA19	47.9843	49.2747	55.3082
TA20	35.2720	36.9672	37.0133
TA21	42.0226	52.8228	**65.0795**
TA22	47.3395	55.5906	**70.5759**
TA23	39.7306	45.1134	49.2781
TA24	37.1046	42.8519	46.2816
TA25	45.9235	47.4199	53.7930
TA26	35.5247	42.0149	43.0694
TA27	42.5357	50.5283	59.3202
TA28	33.7953	41.0114	39.8410
TA29	35.1757	40.9181	42.0036
TA30	37.5511	44.0660	51.4374
TA31	31.4871	34.9636	29.1531
TA32	30.5059	34.2824	28.7970
TA33	37.7007	40.5640	43.0321
TA34	28.0684	32.3588	26.1194
TA35	27.2553	34.3833	30.1136
TA36	28.2665	36.3365	26.8649

5. CONCLUSIONS

Thus, an algorithm for cleaning the data of the general expert population based on:

- competence' checks of respondents;

- identification of patterns (profiles) of respondents' answers to questions, followed by determination of the quality of information in the profile based on the variability of answers, the share of missing data and the meaningfulness of answers.

Conclusions are made and the level of risks of threat of spread of purposeful activity of organized crime in the sphere of fiscal security of Ukraine is determined on the basis of the formed qualitative expert sample.

REFERENCES

[1] Korystin, Oleksandr and Svyrydiuk, Nataliia (2021), "Activities of Illegal Weapons Criminal Component of Hybrid Threats", *Proceedings of the International Conference on Economics, Law and Education Research (ELER 2021)*, vol. 170, 22 March, pp. 86-91. DOI: 10.2991/aebmr.k.210320.016

[2] Svyrydiuk, Nataliia Likhovitskyy, Yaroslav and Polián, Pavel (2021), "Information Threats in the Context of Hybrid War", *Advances in Economics, Business and Management Research. Proceedings of the International Conference on Business, Accounting, Management, Banking, Economic Security and Legal Regulation Research (BAMBEL 2021)*, vol. 188, 27 August 2021, pp. 114-119. DOI:10.2991/aebmr.k.210826.020

[3] EU Serious and Organised Crime Threat Assessment (SOCTA) (2021), Publications Office of the European Union, Luxembourg.

[4] Korystin, Oleksandr and Svyrydiuk, Nataliia (2020), "Methodological principles of risk assessment in law enforcement activity", Nauka i Pravookhorona, vol. 3 (49), pp. 191-198. DOI: 10.36486/np.2020.3(49).19

[5] Korystin, O.Y. Katamadze, G.S. Nekrasov, V.A. Mel'nyk, V.I. and etc. (2021), Fiscal Security of Ukraine – Threats, Risks, Vulnerabilities: Strategic vision, Vidavnichij Dim, Gelvetika, LLC, Kherson, Ukraine.

[6] Basheer M. Al-Maqaleh, Abduhakeem A. Al-Mansoub and Fuad N. Al-Badani (2016), "Forecasting using Artificial Neural Network and Statistics Models", *International Journal of Education and Management Engineering*, vol. 6, no. 3, pp. 20-32. DOI: 10.5815/ijeme.2016.03.03

[7] Junhua, Chen and Fan, Yang (2012), "Research on Audience Rating Statistics of Two-way Digital TV Based on OpenSSL", *IJWMT*, vol. 2, no. 1, pp. 30-37.

[8] Goldammer, P. Annen, H. Stöckli, P.L. and Jonas, K. (2020), "Careless responding in questionnaire measures: Detection, impact, and remedies", *The Leadership Quarterly*, vol. 31 (4): 101384.

[9] Rust, John Kosinski, Michal and Stillwell, David (2021), *Modern psychometrics: the science of psychological assessment*, Milton Park, Abingdon, Oxon; New York, NY: Routledge.

[10] Kovalchuk, T.I. Korystin, O.Y. and Sviridyuk, N.P. (2019), "Hybrid threats in the civil security sector in Ukraine", *Problems of Legality*, vol. 147, pp. 163-175. DOI: 10.21564/2414-990x.147.180550

[11] Bhupesh, Rawat and Sanjay Kumar, Dwivedi, (2019), "Analyzing the Performance of Various Clustering Algorithms", *International Journal of Modern Education and Computer Science*, vol. 11, no. 1, pp. 45-53. DOI: 10.5815/ijmecs.2019.01.06

[12] Ahmed, Fahim, (2018), "Homogeneous Densities Clustering Algorithm", *International Journal of Information Technology and Computer Science*, vol. 10, no. 10, pp.1-10. DOI: 10.5815/ijitcs.2018.10.01

[13] Ahmed, Fahim (2017), "A Clustering Algorithm based on Local Density of Points", *International Journal of Modern Education and Computer Science*, vol. 9, no. 12, pp. 9-16. DOI: 10.5815/ijmecs.2017.12.02

[14] Vy`nogradov, O.G. (2020), "Using the capabilities of the programming language and the R environment in psychological research: A tutorial on basic methods", *Ukrayins`ky`j Psy`xologichny`j Zhurnal, vol.* 2 (14), pp. 28–63.