SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

1/2019
ročník/volume 29

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

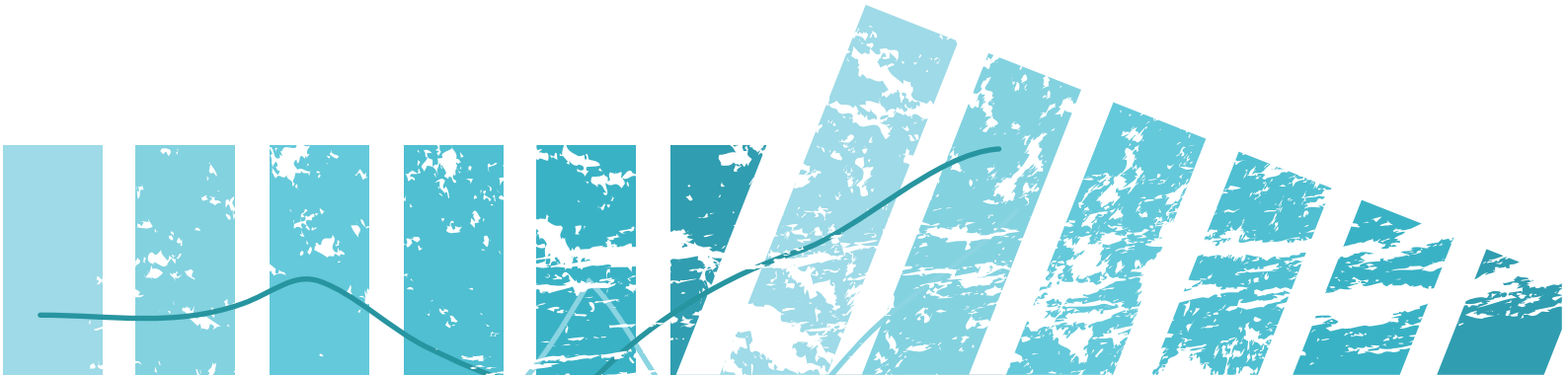
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 3

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 38 – 53

Dátum vydania/Publication date: 15. január 2019/January 15, 2019



Michal PÁLEŠ
Katedra matematiky a aktuárstva, Fakulta hospodárskej informatiky
Ekonomickej univerzity v Bratislave

KVALITA ÚDAJOV A JEJ VÝZNAM PRE AKTUÁROV

DATA QUALITY AND ITS IMPORTANCE – ACTUARIAL VIEW

ABSTRAKT
Možno tvrdiť, že údaje sú v súčasnosti jednoznačne jedným z najhodnotnejších aktív v akejkoľvek sfére podnikania. V poisťovniach dennodenne pribúda množstvo údajov o klientoch, zmluvách, poistných udalostiach, plneniach, majetku, záväzkoch a o iných skutočnostiach, ktoré sa opätovne využívajú v interakcií, v správach a reportoch pre manažment poisťovne a regulárne orgány. Je pochopiteľné že ani v tejto sfére nesmú chýbať pravidlá a samotná kontrola dátovej kvality. Tieto pravidlá určuje aj direktíva Európskej únie Solventnosť II v rámci svojich troch pilieroch výstavby. V príspevku sa budeme zaoberať dátovou kvalitou v kontexte činnosti poisťovne a aktuárskej funkcie. Stručne oboznámime čitateľa s pravidlami, procesom aj manažmentom dátovej kvality. V open-source jazyku R demonštrujeme funkcie a knižnice, ktoré môžu byť využité pre proces čistenia údajov.

ABSTRACT
It can be argued that nowadays data is clearly one of the most valuable assets in any sphere of business. In insurance companies dataabout clients, contracts, insurance events, claims, assets, liabilities and other events that are re-used in interactions and reports for the management of the insurance company and the regulatory bodies are increasing on a daily basis. It is understandable that even in this sphere, the rules and the quality control itself must not be absent. These rules determine the three-pillar structure of the European Union Solvency II Directive. In this paper, we will describe the data quality in the context of the insurance business and the actuarial function. We will briefly present the readers with the rules, process, and data quality management. In the open-source R language, we demonstrate the functions and libraries that can be used for the data-cleaning process.

KĽÚČOVÉ SLOVÁ
Solventnosť II, dátová kvalita, čistenie dát, jazyk R

KEY WORDS
Solvency II, data quality, data cleaning, R language

1. ÚVOD
Informácie hrajú v dnešnom svete veľmi dôležitú rolu. Žijeme v dobe kedy sa informácie stali najdôležitejšou konkurenčnou výhodou. V minulosti ľudia neboli obklopovaní toľkými informáciami ako dnes, a preto dnešná spoločnosť čelí problémom s kvalitou dát (v texte tiež budeme ekvivalentne využívať pojem údaje) nie s ich nedostatkom. Informácie ovplyvňujú každého z nás v našom rozhodovaní a čím sú kvalitnejšie, tak tým nám umožňujú lepšie sa rozhodovať. Kvalitné dáta, z ktorých získame potrebné relevantné informácie nám zabezpečujú jednak znižovanie strát, zvyšovanie ziskovosti, analyzovanie rôznych oblastí, jednoducho povedané vedia nám zefektívniť fungovanie spoločnosti.

Jeden z najdôležitejších aspektov aj projektu Solventnosti II je zabezpečiť vysokú kvalitu dát. Dátová kvalita patrí medzi dôležité otázky, ktoré ovplyvňujú všetkých aktuárov. Či už ide o tvorbu rezerv, oceňovanie, modelovanie alebo vykonávanie iných funkcii. Aktuári vo väčšine prípadov narazia pri vykonávaní svojej profesie na údaje, ktoré sú buď neúplné alebo nepresné. Ďalším dôvodom zavedenia Solventnosti II je, že existujúce požiadavky Solventnosti I predstavujú jednoduchý vzorec, ktorý nie je dostatočne citlivý na riziká. Samozrejme, Solventnosť II nie je jedinou regulačnou výzvou, na ktorú sa musia poisťovne pripraviť. Poisťovne musia zohľadňovať aj navrhované nové medzinárodné účtovné štandardy, napríklad IFRS 17 [16].

V príspevku sa zameriame na dôležité pojmy, ktoré sa týkajú danej problematiky. Stručne budeme analyzovať požiadavky stanovené Solventnosťou II na poisťovacie a zaisťovacie subjekty z hľadiska nárokov na údaje. Budeme definovať dátovú kvalitu z hľadiska praxe, uvedieme aké typy dát sa používajú v poisťovniach a s akými typickými problémami sa poisťovne stretávajú, tiež príčiny a dopady nekvalitných dát, poukážeme na jednotlivé kroky procesu dátovej kvality a predstavíme aj praktickú ukážku v jazyku R.

2. BIG DATA, DATA SCIENCE, DATA MINING, MACHINE LEARNING

Termín Big Data sa používa od deväťdesiatych rokov minulého storočia. Ide o relatívne nový pojem, avšak akt zhromažďovania a ukladania veľkého množstva informácií na prípadné analýzy vznikol skôr.

Ide o pojem, ktorý opisuje veľké množstvo údajov a máme tým na mysli štruktúrované, čiastočne štruktúrované a tiež neštruktúrované údaje, ktoré dennodenne firmy zaplavujú. Dôležité však nie je ani tak množstvo údajov ako to, čo jednotlivé organizácie s nimi zamýšľajú respektíve ako s nimi narábajú. Big Data sa teda stávajú čoraz dôležitejším nástrojom v priebehu rozhodovania o strategických podnikateľských krokoch a vedú k prijímaniu lepších rozhodnutí.

McKinsley definoval Big Data ako údaje, ktorých veľkosť je nad rámec možností typických databázových softvérových nástrojov na zachytenie, spravovanie, ukladanie a analyzovanie. Táto definícia je úmyselne subjektívna a nezahŕňa v sebe aké veľké musia byť dáta, aby sme ich považovali za Big Data.

Laney, priemyselný analytik, tiež prispel k vývoju pojmu Big Data. Napriek tomu, že tento pojem nepoužíval, predpokladal, že spravovanie údajov v elektronickej podobe sa stane dôležitým a náročným, a tak v roku 2001 definoval tzv. 3V model. Neskôr sa k tomuto modelu pridali ešte dve „V“ [8], [15]. Sú to začiatočné písmená anglických pojmov:

- **Data Volume** (*objem údajov*),
- **Data Velocity** (*rýchlosť údajov*),
- **Data Variaty** (*rôznorodosť údajov*),
- **Data Veracity** (*pravosť údajov*),
- **Data Value** (*hodnota údajov*).

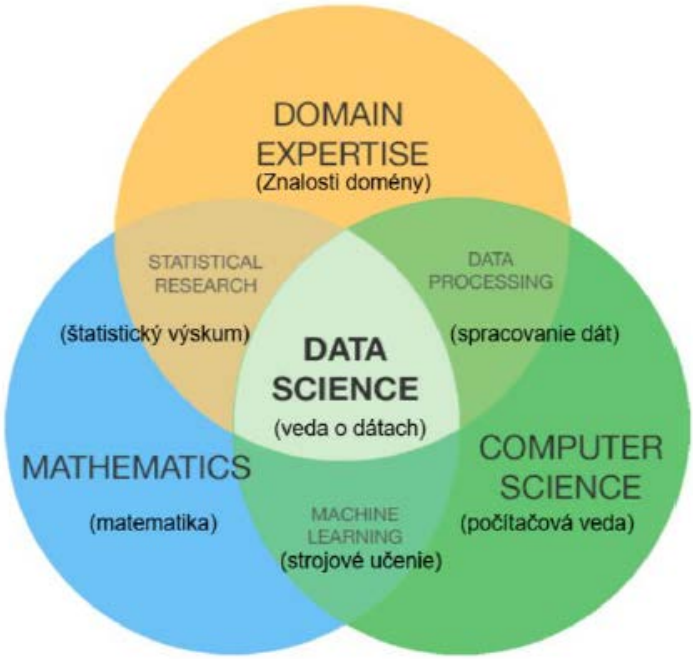
Pre viac informácií o týchto úrovniach pozri napríklad [7].

Stále viac sa objavujú nové výrazy, ktoré súvisia so spracovaním a riadením údajov, potom prepojenie Data Science s ostatnými úrovňami zobrazuje obrázok č. 1.

Proces Data Science je oveľa viac zameraný na technické schopnosti spracovania akýchkoľvek typov údajov. Na rozdiel od Data Mining a Machine Learning (strojové učenie) je zodpovedný za pôsobenie vplyvu údajov v konkrétnom produkte alebo organizácii.

Strojové učenie je podoblasťou umelej inteligencie, zaoberajúce sa algoritmami a technikami, ktoré umožňujú počítačovému systému „učiť sa“. Učením v danom kontexte rozumieme takú zmenu vnútorného stavu systému, ktorá zefektívni schopnosť prispôsobenia sa zmenám okolitého prostredia. Strojové učenie sa značne prelína s oblasťami štatistiky a Data Mining (v jazyku R samostatná oblasť záujmu) a má široké uplatnenie. Jeho techniky sa využívajú napr. v biomedicínskej informatike, rozlíšenie nelegálneho použitia kreditných kariet, rozpoznávanie reči a písaného textu, či mnohé ďalšie. Rozlišujeme rôzne rozdelenie algoritmy, základné druhy úloh aj modely a techniky využívané v rámci strojového učenia.

Obrázok č. 1: Data Science prepojenie



Zdroj: [16]

3. SOLVENTNOSŤ II A KRITÉRIA NA POSÚDENIE KVALITY ÚDAJOV

Európska únia zavádza smernicu Solventnosť II s cieľom zvýšiť ochranu poistencov v EÚ (pozri tiež [2]). Smernica sa vzťahuje na všetky poisťovne, ktoré pôsobia v EÚ a umožňuje tak lepšie pokryť riziká, ktorým poisťovne musia čeliť. Predstavuje nové štandardy riadenia rizík pre spoločnosti s cieľom zaručiť ich prežitie počas ťažkého obdobia, ako sú napríklad povodne, búrky. Podľa nových pravidiel, poisťovne musia držať určité množstvo kapitálu proti rizikám, ktorým sú vystavené. Zatiaľ čo predchádzajúce regulačné požiadavky (Solventnosť I) boli založené predovšetkým na historických dátach, nové nariadenie vyžaduje posúdenie budúceho vývoja, ktorý by mohol ovplyvniť finančnú situáciu poisťovateľa. Štruktúru Solventnosti II rozdeľujeme do troch pilierov, ktoré sú zobrazené na obrázku č. 2 [11].

Aby bolo možné v rámci **I piliera** vypočítať technické rezervy a kapitálovú požiadavku na solventnosť (SCR), poisťovne potrebujú zbierať rôzne údaje z rôznych typov informačných zdrojov, ktoré môžu byť odlišné. Ak poisťovňa

používa interný model na výpočet regulačných kapitálových požiadaviek je nesmierne dôležitá kvalita údajov použitých na validáciu interného modelu. V kontexte s IMAP (vnútorný model schvaľovacieho procesu) sa vyžaduje dokumentácia o vstupných a výstupných údajov, ktoré boli použité v daných modeloch. Preukázanie dátovej kvality je však rovnako dôležité a reguláciou vyžadované aj v prípade aplikácie štandardného vzorca.

Obrázok č. 2: Základná štruktúra (piliere) Solventnosti II

Zdroj: [11]

Čo sa týka dátovej kvality v **pilieri II** poisťovne sa stretávajú s problémami pri riadení operačných rizík. Veľké percento operačných rizík je spôsobené nízkou kvalitou dát počas prevádzky spoločnosti. Napríklad duplicitné platby za škody, nesprávne poistné odhady môžu súvisieť s nízkou kvalitou dát. S cieľom účinne zmierniť tieto riziká, poisťovne potrebujú použiť primerané kontroly na zistenie a prevenciu dátovej kvality.

Úlohou **piliera III** je vybudovať väčšiu mieru transparentnosti, ktorá by posilnila trhový mechanizmus a tiež kontrolu rizika. Tento pilier zahŕňa v sebe požiadavky, ktoré sa týkajú vykazovacej povinnosti voči orgánom dohľadu a zverejňovania informácii. Poisťovne by mali poskytovať periodické reporty o svojich operáciách, ktoré zahŕňajú údaje zosúladené s inými finančnými výkazmi na zvýšenie ich spoľahlivosti. Procesy a systémy použité na generovanie reportov by mali byť dostatočne transparentné, aby mohli sledovať získané údaje až do zdrojového systému.

Piliere II a III smernice Solventnosti II teda predstavujú rozsiahle požiadavky na riadenie a kvalitu dát. Zahŕňa to nie len vytvorenie nových súborov dát a reportov, ale aj štandardy na riadenie dát, ktoré musia byť transparentné a plne kontrolovateľné. EIOPA (zodpovedná za Solventnosť II) vyžaduje, aby bol zavedený *rámec správy dátovej kvality* (*Data Quality Management Framework*) ako súčasť procesu IMAP, ktorý je tiež relevantný pre vlastné posúdenie rizika a solventnosti (ORSA) [10], [11]. Smernica Solventnosť II zdôrazňuje dôležitosť zavedenia štruktúry manažmentu dátovej kvality s cieľom zaručiť neprestajnú a dostatočnú kvalitu dát. Manažment dátovej kvality teda predstavuje nepretržitý proces, ktorý by mal byť rozdelený do štyroch fáz.

Definovanie údajov predstavuje identifikáciu a analýzu informačných tokov vstupujúcich do výpočtov. Tieto údaje musia byť zdokumentované, takisto aj jednotlivé položky by mali byť opísané. Ide o komplexný zoznam údajov, ktoré sa vyžadujú pre príslušné procesy (napr. proces tvorby technických rezerv).

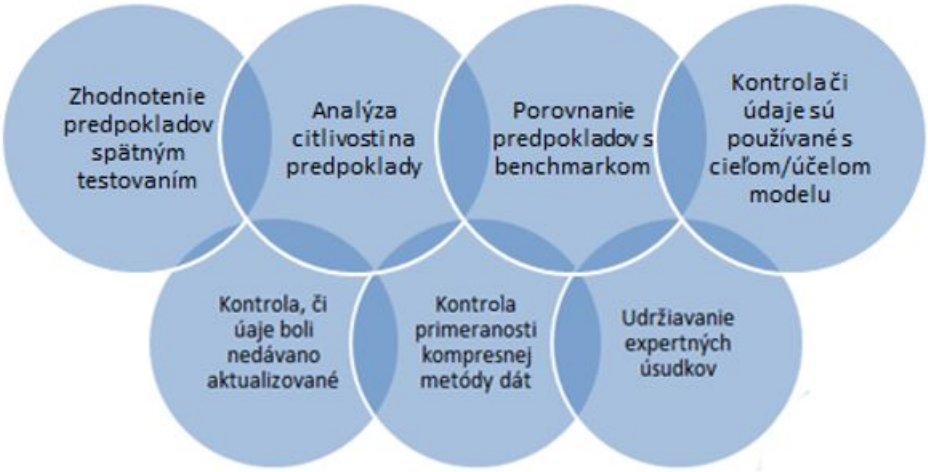
Posúdenie kvality údajov predstavuje overovanie údajov na základe kritérií: úplnosti, presnosti a vhodnosti (pozri ďalej). Najmä vtedy, keď údaje poskytujú tretie strany, by sa mali pri **hodnotení kvality** zohľadniť aj kanály použité na zhromažďovanie, ukladanie, spracovanie a prenos údajov.

V prípade, že poisťovateľ identifikoval problém, mal by sa pokúsiť **vyriešiť daný problém** a zároveň odstrániť jeho nedostatky. Následne všetky vzniknuté problémy a nedostatky by mali byť zdokumentované spolu s návrhom ako zlepšiť danú situáciu. V prípade, že by nastala situácia kedy by sa daný problém nedal vyriešiť, musí byť tiež zdokumentovaný. Poisťovne by sa mali neustále snažiť a pracovať na zlepšovaní svojich interných procesov, aby zabezpečili primeranú kvalitu údajov.

Kvalita údajov by mala byť pravidelne **sledovaná a kontrolovaná** na základe identifikovaných ukazovateľov výkonnosti. Ide predovšetkým o monitorovanie výkonnosti príslušných informačných systémov a kanálov použitých na zhromažďovanie, ukladanie, prenášanie a spracovávanie údajov. Veľmi dôležitý je aj odborný úsudok.

V rámci poisťovní by mali byť zavedené vhodné interné procesy a postupy, ktoré by zabezpečovali takú kvalitu dát pre výpočty, napr. ocenenie technických rezerv, aby vedeli pokryť oblasť manažmentu kvality dát, takisto interné procesy, ktoré sa týkajú identifikácie, zberu a spracovania údajov [5], [11], [18].

Obrázok č. 3: Nástroje na posúdenie vhodnosti dát v poisťovni



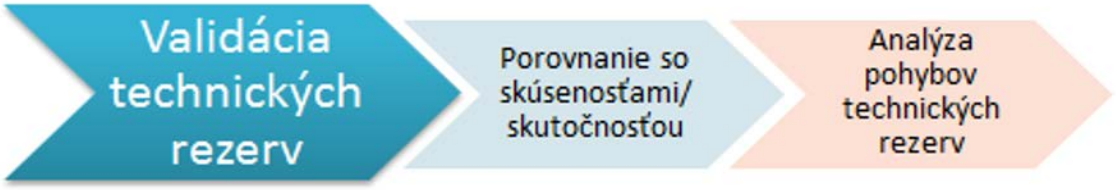
Zdroj: [17]

V rámci smernice Solventnosť II jedným z aspektov je mať k dispozícii správne, vhodné údaje a vedieť zabezpečiť vysokú kvalitu údajov. Solventnosť II definuje tri štandardy pre dátovú kvalitu, podľa článku 82 tejto smernice, a to:

- **presnosť** (*Accuracy*),
- **úplnosť** (*Completeness*),
- **vhodnosť** (*Appropriateness*).

Posúdenie jednotlivých kritérií vhodnosti a úplnosti sa môže vykonávať na úrovni celého portfólia, pričom pri presnosti je lepšie upriamiť pozornosť na každú položku. Taktiež kritéria je potrebné nastaviť tak, aby boli v súlade s princípom proporcionality. Princíp proporcionality by nemal určite viesť k znižovaniu kvality procesu zberu údajov ale má zabezpečovať úplnosť, vhodnosť a hlavne presnosť použitých údajov. Požiadavky sú nastavené na všeobecnej úrovni, pretože zostáva na poisťovniach, ako zabezpečia, aby spĺňali tieto štandardy [4], [5], [11]. Príklad niektorých nástrojov na posúdenie vhodnosti dát v poisťovni zobrazuje obrázok č. 3 a porovnanie so skúsenosťou obrázok č. 4.

Obrázok č. 4: Validácia technických rezerv



modelované výplaty poistných udalostí	1 000 000
skutočné výplaty poistných udalostí	950 000
rozdiel	50 000
5 % rozdiel je v poriadku?	

Zdroj: [17]

Autorka v prezentácii [17] uvádza niektoré postupy ako poisťovňa môže získať komfort s kvalitou dát a čo by malo byť súčasťou dátovej kvality:

- **dokumentácia, ktorá pokrýva kompletný dátový cyklus (*end to end*)**,
- **jasne definované kritériá kvality**,
- **pravidelné posúdenie kvality dát**,
- **definovanie jasného dátového vlastníctva (*data ownership*)**.

Tento komfort možno dosiahnuť:

- **vytvorením zoznamu dát**,
- **vytvorením diagramov toku údajov**,
- **posúdením kvality dát (*presnosť*, *úplnosť*, *vhodnosť*)**,
- **riešením chybovosti v dátach**.

4. PRAVIDLÁ A PROCES DÁTOVEJ KVALITY

V časti 3 sme uviedli kritériá na posúdenie kvality dát, ktoré požaduje Solventnosť II. V praxi sa však využívajú okrem týchto dimenzií aj iné, ktoré uvádza tabuľka č. 1.

Je potrebné poznamenať, že pri spracúvaní a analýze údajov by sa mali používať také údaje, ktoré majú minimálne stanovenú úroveň kvality. Kvalitu dát vždy posudzujeme na základe vopred stanovených požiadaviek používateľa údajov a tiež pre daný zámer ich použitia.

V poisťovniach, v dôsledku Solventnosti II, využívajú tzv. aktuárske, finančné, majetkové a rizikové údaje, ktoré sú kategorizované ako analytické údaje. Na obrázku č. 5 môžeme vidieť jednotlivé typy analytických údajov. Analytické údaje sa svojou povahou líšia od transakčných alebo prevádzkových údajov, ktoré poisťovne

tradične používajú na manipuláciu a skladovanie. Za kľúčové rozdiely môžeme pokladať, že:

- pochádzajú z rôznych zdrojov, predovšetkým z finančných a aktuárskych systémov, ale tiež napríklad z externých systémov správy fondov,
- zvyčajne vyžadujú vyššiu úroveň granularity v údajoch ako sa vyžaduje na reguláciu a súlad účelu vykazovania,
- môžu prejsť komplexnými agregáciami a transformáciami. Údaje musia byť tiež dôkladne zosúladené do jedného zdroja.

Solventnosť II funguje ako katalyzátor, ktorý vedie poisťovateľov k tomu, aby zvážili ako spracúvajú analytické údaje [10].

Tabuľka č. 1: Dimenzia dátovej kvality

DIMENZIA	OPIS
Dostupnosť	*dispozícia a získateľnosť údajov*
Veľkosť a granularita	*veľkosť údajov a ich granularita zodpovedá vykonaným úlohám*
Vierohodnosť	*pravdivosť a dôveryhodnosť údajov*
Úplnosť	*úplnosť, dostatočnosť, rozsiahlosť a detailnosť údajov*
Reprezentácia	*vhodná štruktúra údajov na reprezentáciu*
Konzistentnosť	*rovnaký formát reprezentácie údajov*
Spracovanie	*ľahká spracovateľnosť a použiteľnosť na rôzne úlohy*
Bezchybnosť	*presnosť a hodnovernosť údajov*
Interpretovateľnosť	*jasná definícia informácií, zodpovedajúci jazyk, jednotky, správne symboly*
Objektivita	*nestrannosť a nezaujatosť informácií*
Relevantnosť	*použiteľnosť informácií a ich užitočnosť pre vykonávané úlohy*
Reputácia	*spoľahlivosť informácií v súvislosti s ich zdrojom alebo obsahom*
Bezpečnosť	*bezpečnostné pravidlá prístupu k informáciám*
Včasnosť	*včasná dostupnosť informácií pre vykonávané úlohy*
Zrozumiteľnosť	*ľahká pochopiteľnosť a zrozumiteľnosť informácií*
Pridaná hodnota	*prínos informácií a výhody ich použitia*

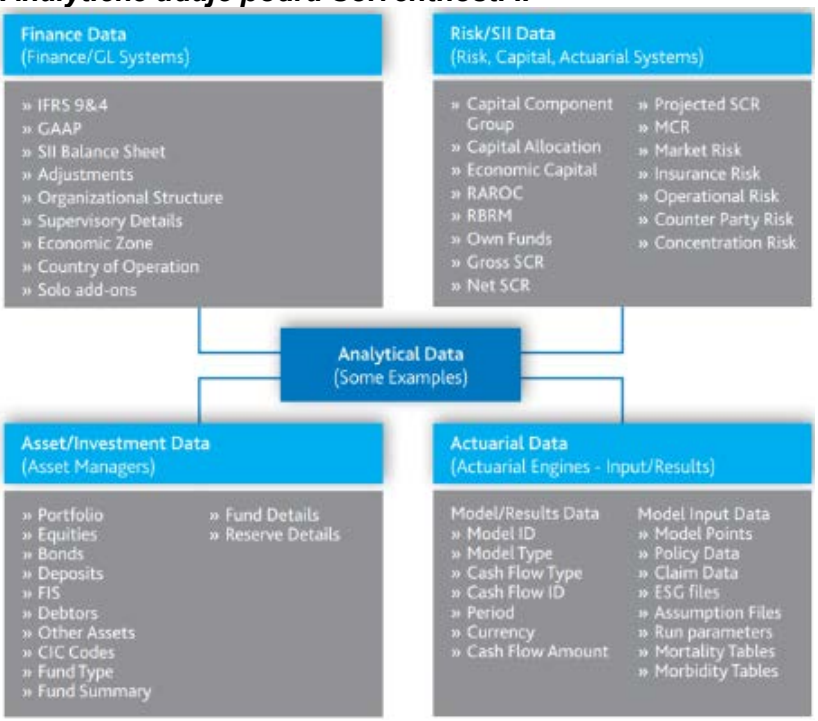
Zdroj: spracované podľa [14], [16]

Aktuárske modelovanie je tradične doménou desktopových modelovacích systémov. Mať homogénne údaje je bezvýznamné, pokiaľ ich nemožno agregovať spôsobom, ktorý ich podporuje. Solventnosť II vyžaduje agregované zobrazenie viacerých množín a surové údaje môžu vyžadovať sofistikované analytické metódy. Aby sa zabezpečilo, správne zoskupenie údajov musia sa štandardy aplikovať na zber a analýzu údajov. Existuje veľa potenciálnych vstupov do aktuárskych modelov, napríklad konštrukcia úmrtnostných tabuliek (viac v [16]).

Poisťovatelia sa často stretávajú s nedostatočnou kvalitou údajov, ktorá je vo väčšej miere spôsobená zlou interpretáciou údajov. Zlá interpretácia zas vedie k zlým rozhodnutiam a tiež k nedorozumeniam. Jednoducho povedané informácie, ktoré získame sú neúplné, skreslené alebo úplne chybné, teda nemusia odrážať

skutočnosť reálneho sveta. Môže to byť spôsobené množstvom neštruktúrovaných údajov uložených v mnohých systémoch. Viaceré systémy sú už zastarané a iné sú založené na aktuárskych softvéroch. Hlavný problém spočíva v nových aplikáciách, ktoré boli pridané do starších softvérov a vytvárajú tak viacvrstvové a potenciálne redundantné IT architektúry.

Obrázok č. 5: Analytické údaje podľa Solventnosti II



Zdroj: [10]

Zdroje údajov nie sú vždy známe, a preto pri vzniku rôznych problémov s údajmi sa častokrát poisťovne stretávajú s tým, že nevedia na koho sa obrátiť, vzhľadom k tomu, že nie sú určení vlastníci zdrojových údajov. Používanie rôznych zdrojov tiež spôsobuje problémy s číslami, ktoré by mali byť v rôznych reportoch rovnaké, no sú iné a dôvod nie je známy.

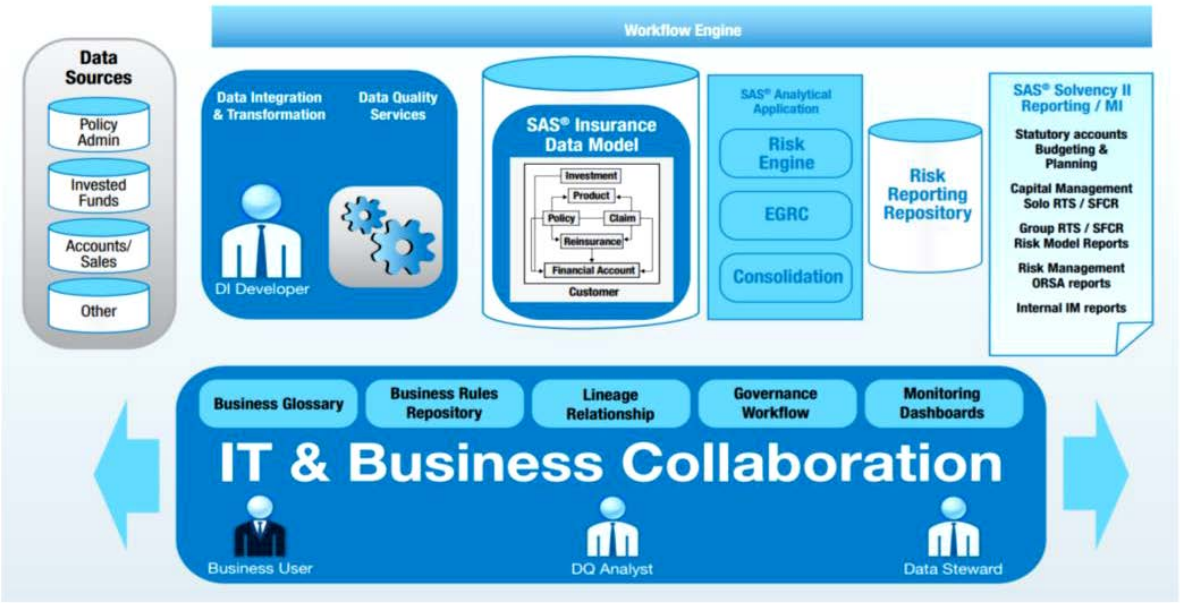
Ďalším problémom je, že požiadavka Solventnosti II má za úlohu zabezpečiť, aby všetky údaje boli presné, úplne a vhodné a stanovuje tiež normy pre dátovú kvalitu. Problém však spočíva v tom, že poisťovne nie sú vždy schopné definovať uvedené tri kritériá (presnosť, úplnosť, vhodnosť), napríklad keď sa jedná o katastrofické riziká.

Vo všeobecnosti však platí, že v každom systéme sa vyskytujú chyby v údajoch ale spoločnosti by sa mali snažiť, aby ich tam bolo čo najmenej. Sú dve základné príčiny prečo sú údaje nekvalitné. Prvým dôvodom je veľký výskyt chýb, ktoré sú do určitej miery spôsobené zlyhaním ľudského faktora. Môže teda dôjsť k chybovosti: preklep, nesprávny formát, nesprávne pole a mnohé iné. Druhým dôvodom je nekonzistencia informačných systémov podniku. Znamená to, že informačné systémy sa skladajú z rôznych systémov, ktoré sú založené na rôznych technológiách a tieto systémy implementujú rôzni dodávatelia. S tým môže súvisieť zastaranosť údajov. V databázach je zachovaná určitá statickosť reálneho sveta, ktorá sa samozrejme mení. Nie vždy sa však nové zmeny dostanú do všetkých systémov a to spôsobuje, že údaje sa kazia a nemusia odrážať skutočnosť reálneho sveta. Môže ísť aj o legislatívne problémy. Štát nariaďuje, akým spôsobom

majú byť uchovávané údaje o klientoch a v prípade zlého spracovania môžu byť spoločnosti sankcionované.

Dopady nekvalitných dát môžu byť spôsobené jednak výskytom duplicitných záznamov, ako aj nekonzistentnosťou, nepresnosťou a zastaranosťou údajov. To všetko spôsobuje chyby v reportoch, ktoré vedú k zlým rozhodnutiam a premeškaným príležitostiam. Nepresné, neaktuálne a neúplne údaje sú zlé pre akýkoľvek biznis, najmä pokiaľ ide o ziskovosť a konkurenčnú výhodu. Nesprávne údaje môžu viesť aj k strate z príjmov. Napríklad v oblasti poistenia, zlé údaje môžu spôsobiť stratu príjmov z poistného, ak poistné bolo nastavené príliš nízko. Na to, aby sme vedeli čeliť týmto dôsledkom nekvalitných dát, musíme najskôr poznať dôvody predchádzajúce vzniku nekvalitných údajov a vďaka tomu vieme prijať lepšie opatrenia [1], [6], [9].

Obrázok č. 6: SAS dátový manažment pre Solventnosť II



Zdroj: [15]

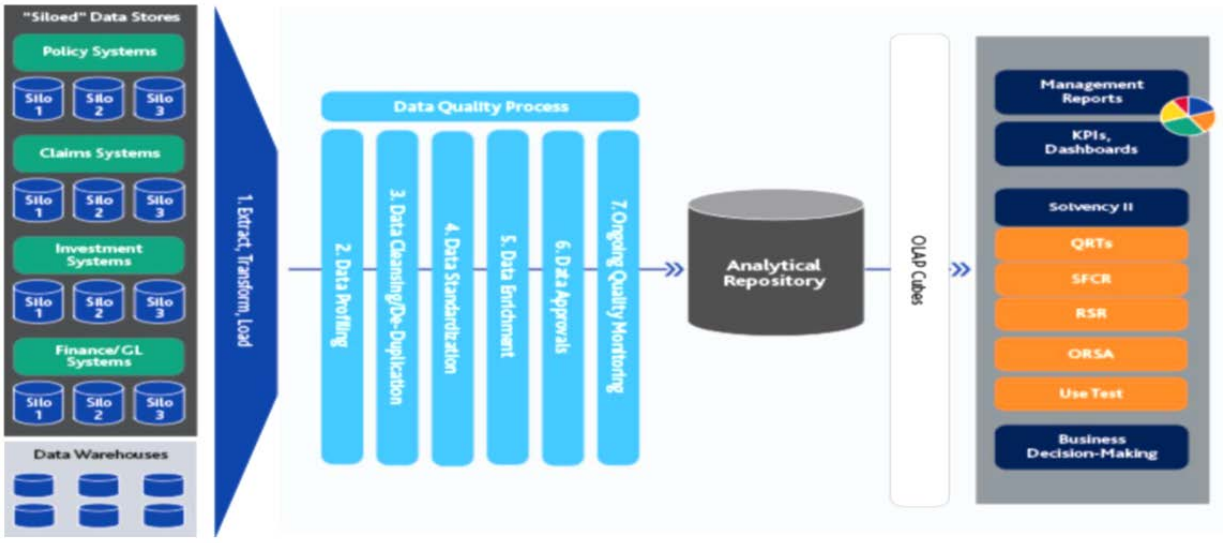
V záujme ďalšieho zlepšovania presnosti údajov poisťovatelia by mali spĺňať pravidlá dátovej kvality. Rôzni predajcovia ponúkajú nástroje na dátovú kvalitu, ktorá môže zahŕňať tisícky pravidiel. Tieto zahŕňajú niekoľko všeobecných pravidiel spolu s niektorými špecifickými pravidlami poisťovacieho priemyslu. Okrem toho takéto nástroje umožňujú poisťovateľom definovať doplnkové pravidlá, ktoré sú špecifické pre ich podnikanie [10]. Obrázok č. 6 napríklad popisuje komerčný SAS dátový manažment, ktorý je určený pre poisťovne (v zmysle Solventnosti II). Zlepšenie dátovej kvality je multilaterálny proces. Podľa Solventnosti II sú údaje považované za vysokokvalitné, ak sú „*Fit for Purpose*“ z hľadiska ich účelu použitia.

Možnosťami, ako jednotlivé poisťovne (ale aj iné organizácie) postupujú pri procese dátovej kvality sa zaoberá [16]. Obrázok č. 7 znázorňuje dátový reťazec medzi zdrojovými systémami, reportovaním a poukazuje aj na proces dátovej kvality, ktorý pozostáva z nasledujúcich krokov:

- **uchovanie dát v dátových skladoch** (*Data Stores*),
- **získavanie, transformovanie, načítanie** (*Extraction, Transformation, Floating),*
- **profilovanie dát** (*Profiling Data*),

- **čistenie dát** (*Cleansing Data*),
- **dátová štandardizácia** (*Data Standardization*),
- **obohacovanie dát** (*Data Enrichment*),
- **schvaľovanie dát** (*Data Approvals*),
- **monitorovanie dát** (*Monitoring*),
- **ukladanie do analytického úložiska** (*Analytical Repository*),
- **reportovanie** (*Reporting*),
- **výstupy** (*Outputs*).

Obrázok č. 7: Proces dátovej kvality pre Solventnosť II



Zdroj: [10]

5. VYUŽITIE JAZYKA R NA ČISTENIE ÚDAJOV

Po úspešnom extrahovaní, transformácii a načítaní údajov zo zdrojových systémov sa údaje uložia do dátového skladu a nasleduje ďalšia etapa – čistenie. Čistenie údajov sa skladá najmä z dvoch krokov:

1. **identifikovanie chýb**,
2. **opravovanie chýb**.

Čistenie je najdôležitejšia etapa, pretože zaisťuje dátovú kvalitu v dátovom sklade. Wickham sa podrobne venoval čisteniu údajov a vo svojej práci (*Tidy Data* – čisté údaje) popisuje päť najčastejších problémov s tzv. *Messy* (špinavé) dátovými súbormi [19]:

- **hlavičky stĺpcov sú hodnoty, nie názvy premenných**,
- **v jednom stĺpci sa ukladajú viaceré premenné**,
- **premenné sú uložené v riadkoch a stĺpcoch,**
- **v rovnakej tabuľke sú uložené viaceré typy pozorovacích jednotiek,**
- **jedna pozorovacia jednotka je uložená vo viacerých tabuľkách**.

Na ukážku manipulácie s údajmi v tejto fáze využívame programovací jazyk R, kde možno použiť knižnice ako tidyr, tidyverse, eeptools, dplyr, ggplot2 a i. Okrem spomínaných knižníc sa pri čistení dát, respektíve manipulácii s dátami využívajú tiež: Reshape2, Janitor alebo Lubridate. Podrobnejší opis k jednotlivým knižniciam možno získať z https://cran.r-project.org/web/packages/. Prácu s jazykom R v oblasti aktuárstva podrobne popisuje učebnica [13] a tiež [12].

Na prácu možno využiť aj iný programovací jazyk, ako je napríklad Python, ktorý ponúka tiež knižnice na manipuláciu a čistenie dát. Jedná sa napríklad o knižnice Pandas, NumPy a iné. Je len na používateľovi, ktorou cestou sa dopracuje k požadovanému výsledku. Poisťovne zvyknú pri svojich projektoch využívať služby GitHub, ktoré im uľahčujú plánovanie, vytváranie, manažovanie projektov a mnoho iného v oblasti analýzy dát v rámci spoločnosti.

Nižšie si teda ukážeme zaujímavý príklad čistenia údajov, keď najskôr vytvoríme tréningový dataset čistých (*tidy*) údajov, potom z nich urobíme údaje, ktoré nie sú čisté (*messy*) a následne sa pokúsime rôznymi technikami spätne získať pôvodnú databázu údajov. Navrhovaný postup je tiež využitý používateľmi jazyka R na https://r-bloggers.com/. Opis riešenia je priamo uvedený v komentároch a funkciách kódu jazyka R s príslušným výstupom.

```
library(formattable)
library(plyr)
library(dplyr)
library(tidyr)
library(stringr)

# vytvorenie „čistého“ datasetu

var1_text = c("Sachin", "Sourav", "Rahul", "Laxman")
var2_text = c("Virat", "Jinx", "Pujara", "Rohit")
sep1 = ":"
sep2 = "|"
no_rows = 100
set.seed(9653)
d1 = data.frame(id = 1:no_rows,
            retired = sample(x = var1_text, size = 10,
replace = TRUE),
            current = sample(x = var2_text, size = 10,
replace = TRUE),
            garbage = paste0("my_var", 1:no_rows),
            stringsAsFactors = FALSE)
formattable(head(d1))
```

id	retired	current	garbage
1	Sachin	Pujara	my_var1
2	Rahul	Virat	my_var2
3	Laxman	Virat	my_var3
4	Sourav	Virat	my_var4
5	Sachin	Virat	my_var5
6	Sachin	Virat	my_var6

```
# umelé „znečistenie" datasetu (príprava na ukážku)

d2 = d1
var_names = names(d1)[-1]
d2$var1_pair = paste(var_names[12], d2$retired, sep = sep1)
d2$var2_pair = paste(var_names[12], d2$current, sep = sep1)
d2$var3_pair = paste(var_names[13], d2$garbage, sep = sep1)
d2 = d2[, c("id", "var1_pair", "var2_pair", "var3_pair")]
d3 = d2

d3$text = NA
d3$text[4 * (1:25) - 3] = paste(d3$var1_pair[4 * (1:25) - 3],
                                d3$var2_pair[4 * (1:25) - 3],
                                d3$var3_pair[4 * (1:25) - 3],
                                sep = sep2)
d3$text[4 * (1:25) - 2] = paste(d3$var2_pair[4 * (1:25) - 2],
                                d3$var3_pair[4 * (1:25) - 2],
                                sep = sep2)
d3$text[4 * (1:25) - 1] = paste(d3$var3_pair[4 * (1:25) - 1],
                                d3$var2_pair[4 * (1:25) - 1],
                                d3$var1_pair[4 * (1:25) - 1],
                                sep = sep2)
d3$text[4 * (1:25)] = d3$var2_pair[4 * (1:25)]

d3 = d3[ , c("id", "text")]
formattable(head(d3))
```

id	text
1	retired:Sachin|current:Pujara|garbage:my_var1
2	current:Virat|garbage:my_var2
3	garbage:my_var3|current:Virat|retired:Laxman
4	current:Virat
5	retired:Sachin|current:Virat|garbage:my_var5
6	current:Virat|garbage:my_var6

```
# príklad čistenia údajov

len = max(str_count(string = d3$text, pattern =
paste0("[",sep2,"]")))
vec_names = paste0("X", 1:(len + 1))

d2_rev = d3 %>%
  separate(col = "text", into = vec_names, sep =
paste0("[",sep2,"]"), extra = "drop")
d3_rev = d2_rev %>%
  gather(key = "temp_var", value = "kv_pair", -id, na.rm =
TRUE) %>%
```

```
  select(-temp_var) %>%
  separate(col = "kv_pair", into = c("key", "val"), sep =
paste0("[",sep1,"]"), extra = "drop") %>%
  spread(key = "key", value = "val")
```

Záverečný výstup je potom podobný *tidy* dataset-u, ktorý sme vygenerovali na začiatku s výnimkou poradia stĺpcov a niektorých hodnôt *NA*. Tieto úpravy možno realizovať využitím ďalších funkcií.

id	current	garbage	retired
1	Pujara	my_var1	Sachin
2	Virat	my_var2	NA
3	Virat	my_var3	Laxman
4	Virat	NA	NA
5	Virat	my_var5	Sachin
6	Virat	my_var6	NA

6. ZÁVER

V roku 2016 vstúpila do platnosti smernica Solventnosť II, t. j. metodický rámec na reguláciu poisťovní, ktorý predstavuje systematický prístup k riadeniu rizík, vedie k lepšiemu ohodnocovaniu rizík a tým k záruke vyššej ochrany poistených osôb. Solventnosť II taktiež poskytuje viacero intencií ako posudzovať a zabezpečiť kvalitu údajov, ktorá je z pohľadu vykazovania a reportovania taká dôležitá pre regulačné orgány i pre samotný manažment poisťovne. Túto dôležitosť dokumentujú aj prednášky v národných aktuárskych spoločnostiach (napr. Slovenská spoločnosť aktuárov (aktuar.sk), Česká společnost aktuárů (actuaria.cz), európskych i svetových aktuárskych asociácií (AAE, IAA), poradenských a konzultačných spoločností (De-loitte, KPMG, ...), spoločností pre vývoj softvéru (SAS,...).

Príspevok poukazuje na základné kritériá, ktoré sú kladené na dátovú kvalitu – a to presnosť, úplnosť a vhodnosť – a vysvetľuje troj-pilierovú štruktúru Solventnosti II z pohľadu kvality údajov. Nakoľko dátová kvalita predstavuje nepretržitý proces, Solventnosť II zdôrazňuje zavedenie štruktúry manažmentu dátovej kvality. Tento proces je rozdelený do štyroch základných fáz a to: definovanie údajov, ohodnotenie dátovej kvality, riešenie problému a nakoniec monitorovanie dátovej kvality. Aktuár by mal teda, okrem iného, venovať zvýšenú pozornosť analýze dátovej kvality a vedieť odpovedať na kľúčové otázky: prečo je dôležitá dátová kvalita pre jeho činnosť, aké údaje používa a aké príčiny a dopady môžu vyvolať nekvalitné údaje.

Proces dátovej kvality môže pozostávať z niekoľkých krokov: extrakcie, transformácie, načítania údajov, profilovania, čistenia, dátovej štandardizácie, obohacovania, schvaľovania a monitorovania dát. V praktickej časti sme ukázali na jednoduchom príklade využitie funkcií open-source jazyka R na manipuláciu s údajmi vo fáze „čistenia“ – presnejšie ak máme v databáze nekonzistentné údaje, ktoré potrebujeme spracovať. Jednoducho povedané na každú analýzu by sme mali svoje údaje upraviť tak, aby sme ich mohli uložiť do predvoleného, ale

univerzálneho formátu a transformovať ich na modelovanie a vizualizáciu pomocou nástrojov, ktoré pracujú s údajmi v tomto formáte. Optimálna štruktúra je, keď každá premenná tvorí stĺpec, každé pozorovanie tvorí riadok, každý súbor údajov obsahuje informácie o jedinom pozorovaní a pod. Týmito predpokladmi sa často vyznačujú veľkoformátové alebo panelové údaje, avšak reálne dáta majú často oveľa komplikovanejšiu štruktúru (napr. jeden stĺpec obsahuje údaje o viacerých premenných namiesto jednej premennej, názvy stĺpcov predstavujú hodnoty údajov namiesto názvov premenných atď.), a preto si vyžadujú aplikáciu sofistikovanejších techník. Pre procedúru *Data Cleaning* (ktorá sa často uvádza ako súčasť *Data Wrangling* [*Munging*]) sa v jazyku R najčastejšie využívajú knižnice `dplyr`, `tidyr`, `tidyverse`, `eeptools` a i.

Pre ďalšie informácie k tejto aktuálnej problematike 21. storočia, ktorú môžeme posudzovať z rôznych hľadísk (IT, programovacej, aktuárskej, účtovníckej, audítorskej,...) odporúčame čitateľovi využiť rôzne ďalšie dostupné zdroje.

Príspevok bol spracovaný v rámci projektov: VEGA č. 1/1020/18, VEGA č. 1/0647/19 a KEGA č. 021EU-4/2019.

LITERATÚRA

[1] Carnegie Mellon University: Software Engineering Institute [online]. [cit. 21.3.2018] Dostupné na: <https://www.sei.cmu.edu/measurement/research/upload/Loshin.pdf>

[2] CIPRA, T.: Riziko ve financích a pojišťovnictví: Basel III a Solvency II. Praha: Ekopress, 2015. ISBN 978-80-87865-24-8.

[3] Data Science Central: Difference of Data Science, Machine Learning and Data Mining. [online] [cit. 20.3.2018] Dostupné na: <https://www.datasciencecentral.com/profiles/blogs/difference-of-datascience-machine-learning-and-data-mining>

[4] Deloitte: Datová kvalita nejen pro Solvency II. [online]. [cit. 14.10.2017] Dostupné na: <http://www.actuaria.cz/ upload/SAV_ DQ_Petr_Dvorak_PDF.pdf>

[5] EIOPA: Solvency II – Regulatory framework. [online] [cit. 20.2.2018] Dostupné na: <https://eiopa.europa.eu/ regulation-supervision/insurance/solvency-ii>

[6] Forbes: The Importance of Data Quality – Good, Bad or Ugly. [online]. [cit. 22.3.2018] Dostupné na: <https://www.forbes.com/sites/forbesinsights/2017/06/05/the-importance-of-data-quality-good-bad-or-ugly/#1eb028fd10c4>

[7] Linkedin: Big Data – The 5 Vs Everyone Must Know. [online] [cit. 20.3.2018] Dostupné na: <https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know>

[8] McKinsey Global Institute: Big data – The next frontier for innovation, competition, and productivity. [online]. [cit. 20.2.2018] Dostupné na: <https://www.mckinsey.com/~/media/McKinsey/.../MGI_big_data_exec_summary.ashx>

[9] Melissa Global Intelligence: The Impact of Poor Quality Data. [online]. [cit. 21.3.2018] Dostupné na: <http://www.melissadata.com/enews/dataadvisor/articles/ 062011/1.htm>

[10] Moody´s analytics: Analytical Data – How Insurers can improve quality. [online] [cit. 20.3.2018] Dostupné na: <https://www.moodysanalytics.com/-/media/whitepaper/2013/2013-17-07-analytical-data-how-insurers-canimprove-quality.pdf>

[11] Národná Banka Slovenska: Solventnosť II. [online]. [cit. 14.10.2017] Dostupné na:<http://www.nbs.sk/_img/Documents/_Dohlad/ORM/Poistovnictvo/Solventnost_II.pdf>
[12] PÁLEŠ, M.: Aktuárstvo v režime Solventnosť II (S riešenými príkladmi v jazyku R). Bratislava: Vydavateľstvo Ekonóm, 2016. ISBN 978-80-225-4288-3.
[13] PÁLEŠ, M.: Jazyk R v aktuárskych analýzach. Bratislava: Vydavateľstvo EKONÓM, 2017. ISBN 978-80-225-4331-6.
[14] Profinit: Dátová kvalita. [online] [cit. 20.2.2018] Dostupné na: <https://profinit.eu/wp-content/uploads/2015/ 12/02_Datova_kvalita.pdf>
[15] SAS: Big Data – What i tis and why it matters. [online] [cit. 21.2.2018] Dostupné na: <https://www.sas.com/ en_us/insights/big-data/what-is-big-data.html>
[16] SLOVÁKOVÁ, M.: Kvalita údajov a Solventnosť II. diplomová práca: Ekonomická univerzita v Bratislave, 2018.
[17] Slovenská spoločnosť aktuárov: Knošková, N.: Validácia technických rezerv podľa požiadaviek Solventnosti II. Bratislava, 2018.
[18] Universiteit Leiden: Data Quality Management and Solvency II Perspective. [online]. [cit.14.10.20178] Dostupné na: <http://liacs.leidenuniv.nl/assets/Masterscripties/ Altinay-Soyer.pdf>
[19] WICKHAM, H.: Tidy Data. [online] [cit. 10.2.2018] Dostupné na: <http://vita.had.co.nz/papers/tidy-data.pdf>

RESUME

For any business on the market, it is very important to keep the structure of the necessary data because statistical analyses have been carried out on the basis of these data, upon which important decisions are subsequently taken . At present, one of the most important tasks in an insurance company is the data quality management, the necessity of which is to ensure that the data used in the organization is accurate, reliable and fault-free, as laid down in the European Union Solvency II Directive. Insurance companies need to collect different data from various types of information sources for their calculations, which can differ. Similarly, a high percentage of operational risks is due to poor data quality. In order to effectively mitigate these risks, insurance companies need to use adequate checks for the detection and prevention of data quality. The task is also to build a higher degree of transparency that would strengthen the market mechanism as well as risk control. Insurance companies should provide periodic reports on their operations that include data consistent with other financial statements to enhance the reliability of the reports. Processes and systems used for reports generation should be sufficiently transparent to be able to track the data obtained up to the source system. Thus, Solvency II's pillars represent extensive data management and quality requirements. This includes not only the creation of new data sets and reports, but also data management standards, which must be transparent and fully controllable. It is necessary to implement data quality management. By using R language, the analyst can successfully work with a large data database. Several available libraries allow to execute processes for assessment, sorting, processing, cleaning, and subsequent evaluation of data for further needs in the data quality management processes.

PROFESIJNÝ ŽIVOTOPIS

Ing. Michal Páleš, PhD., *od roku 2012 pôsobí ako odborný asistent a tajomník Katedry matematiky a aktuárstva Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave. V rámci pedagogickej činnosti vyučuje predmety matematika, teória pravdepodobnosti, softvérové aplikácie pre aktuárov, teória rizika v poistení, úvod do aktuárstva a vybrané kapitoly z matematiky. Vo svojej vedeckej práci sa orientuje na aktuársku vedu, využitie kvantitatívnych metód v ekonómii a softvérovú podporu riadenia rizík. Je autorom viacerých ocenených vysokoškolských učebníc a vedeckých článkov z oblasti aktuárstva.*

KONTAKT

pales.euba@gmail.com